BEYOND VISUAL ARTIFACTS: KNOWLEDGE-INTENSIVE DETECTION OF AI-GENERATED IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Detectors that rely on pixel-level artifacts or semantic visual flaws can struggle with increasingly realistic AI-generated images (AIGIs). Prior AIGI detectors have not systematically exploited a complementary signal: knowledge inconsistencies between implicit visual claims and real-world facts. Such inconsistencies can arise from generator hallucinations or deliberate creator choices. To evaluate this signal, we introduce **K-AIGI-Bench**, to our knowledge the first benchmark specifically designed for *knowledge-intensive AIGI detection*. It contains carefully human-curated, high-quality images produced by a state-of-the-art image generator, each with a deliberately embedded knowledge inconsistency, alongside topic-matched real images collected from the Internet. To address this task, we further propose **K-VERA** (*Knowledge-grounded Visual Evidence Retrieval and Adjudication*), a training-free, retrieval-augmented multi-agent framework. It identifies the scene and entities, turns implicit visual claims into independent verification tasks, and retrieves webpages and reference images to verify them. It then checks source support and resolves conflicting findings to produce an evidence-grounded verdict. Experiments on K-AIGI-Bench show that K-VERA outperforms all evaluated dedicated AIGI detectors by **over 15%** in balanced accuracy, while producing higher-quality, evidence-grounded explanations. Combining K-VERA with a pixel-level detector further improves performance on three in-the-wild benchmarks. These results establish knowledge-intensive detection as an important enhancement for detecting increasingly realistic AIGIs.

1 INTRODUCTION

Modern image generators produce photorealistic content with unprecedented fidelity (Rombach et al., 2022; Esser et al., 2024; Black Forest Labs, 2024; Google, 2025; OpenAI, 2026c), substantially lowering the barrier to creating deceptive media and posing growing challenges to public trust. Existing AI-generated image (AIGI) detection predominantly relies on *image-intrinsic cues*: either *pixel-level artifacts* (e.g., frequency anomalies, generator fingerprints) (Wang et al., 2020; Ojha et al., 2023; Wang et al., 2023; Chen et al., 2024), which generalize poorly to unseen generators and compression, or *semantic visual flaws* (e.g., unrecognizable text, extra fingers) parsed by multimodal large language models (MLLMs) (Wen et al., 2025; Huang et al., 2025b; Gao et al., 2025).

However, as generative models rapidly reduce both pixel- and semantic-level flaws, purely image-intrinsic inspection falls short. A visually flawless image can still conflict with reality by depicting a nonexistent event, placing a landmark in an impossible layout, or rendering a real device with incorrect components. These errors may arise from generator hallucination or be introduced deliberately by the image creator. Detecting them requires evaluating *knowledge inconsistencies*—discrepancies between implicit visual claims and external world knowledge (Figure 1, left). We refer to this problem as *knowledge-intensive AIGI detection*, which inherently necessitates general, specialized, or time-sensitive factual verification.

To systematically evaluate this setting, we build **K-AIGI-Bench**, to our knowledge the first benchmark specifically designed for *knowledge-intensive AIGI detection*. It contains about 1,000 images comprising around 500 GPT-Image-2 images with deliberately embedded knowledge inconsistencies across eight topics, paired with around 500 topic-matched real images. Each synthetic image is verified to be free of conspicuous low-level artifacts and is annotated with its inconsistency type, a

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

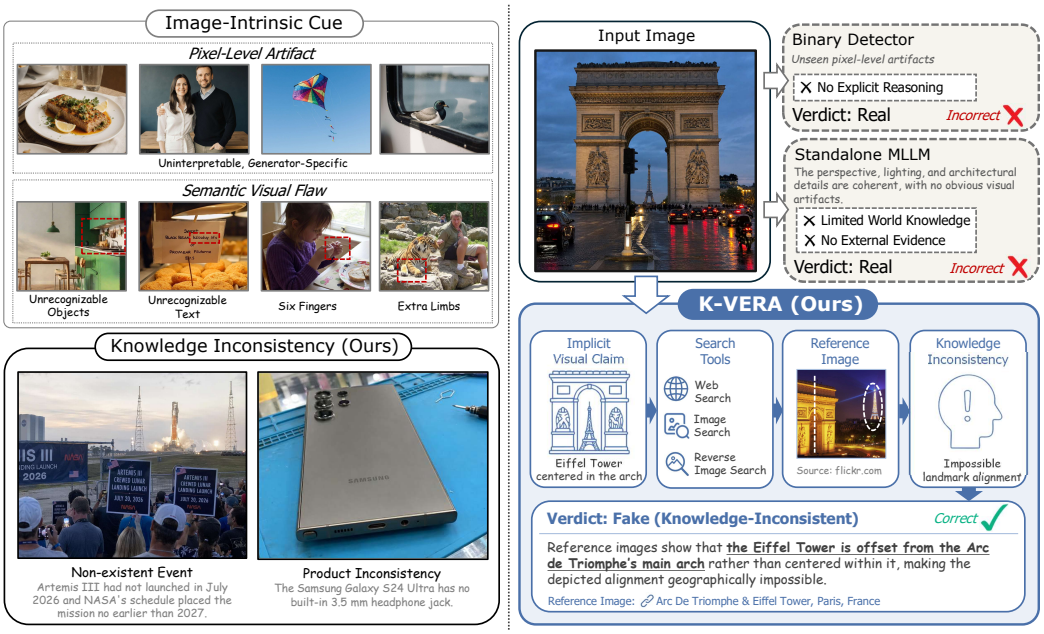

Figure 1: Overview of **two types of AIGI detection signals** (left) and our **K-VERA framework** (right). Unlike image-intrinsic cues (pixel artifacts or semantic flaws), **knowledge inconsistencies** stem from conflicts between implicit visual claims and real-world facts. Where standalone detectors fail due to the lack of detectable flaws, K-VERA retrieves external web evidence to verify implicit visual claims and adjudicates the image as fake.

ground-truth defect explanation, and the required knowledge date, indicating how recent the knowledge must be for verification.

To tackle this challenge, we introduce **K-VERA** (*Knowledge-grounded Visual Evidence Retrieval and Adjudication*), a training-free, retrieval-augmented multi-agent framework. K-VERA identifies visual entities, translates them into targeted verification tasks, retrieves multimodal web evidence, and adjudicates a final real/fake verdict supported by external reference links and images (see the right side of Figure 1).

Our main contributions are summarized as follows:

- We formulate *knowledge-intensive AIGI detection*, a critical yet overlooked setting targeting visually plausible synthetic images whose detection largely depends on identifying conflicts between implicit visual claims and real-world facts.
- We introduce **K-AIGI-Bench**, **the first benchmark** dedicated to this task, containing $\sim$1k images across eight knowledge topics, fully annotated with ground-truth defect explanations and required knowledge dates.
- We propose **K-VERA**, a training-free, retrieval-augmented multi-agent framework that turns implicit visual claims into independent verification tasks and adjudicates retrieved multimodal web evidence. It outperforms the best evaluated dedicated AIGI detector by **17.4%**, and improves performance when combined with a pixel-level detector on three in-the-wild benchmarks.

2 Related Work

2.1 Conventional AIGI Detection

Conventional AIGI detectors classify images using image-intrinsic cues, from learned generator artifacts (Wang et al., 2020) to frequency and neighboring-pixel statistics, pretrained representations, and diffusion reconstruction errors (Frank et al., 2020; Tan et al., 2024; Ojha et al., 2023; Wang

et al., 2023; Chen et al., 2024). Recent methods further enhance cross-generator generalization, for example through semantic-frequency feature fusion, residual feature subspace adaptation, and alignment of real and reconstructed data (Yan et al., 2025a;b; Chen et al., 2025a). Despite these advances, their evidence remains primarily image-intrinsic.

2.2 Explainable and Agentic AIGI Detection

Beyond binary labels, explainable AIGI detectors leverage MLLMs for reasoned decisions and human-understandable explanations. While general MLLMs exhibit zero-shot forensic capabilities (Jia et al., 2024; Li et al., 2025a), subsequent works specialize MLLMs via instruction tuning on forensic datasets (Wen et al., 2025; Li et al., 2025b; Tan et al., 2026), attach spatial localization heads (Xu et al., 2025; Kang et al., 2025; Huang et al., 2025b), incorporate auxiliary forensic vision experts (Zhou et al., 2025; Gao et al., 2025; Chen et al., 2025b), or apply reinforcement learning to refine reasoning (Huang et al., 2025c; Ji et al., 2026; Huang et al., 2025a). Recent agentic systems further employ multi-expert routing (Huang et al., 2026), heterogeneous detector fusion (Yu et al., 2026), dynamic scheduling (Zhu et al., 2026), and multi-agent debate (Liang et al., 2025). Nonetheless, these systems center on image-intrinsic artifacts, failing to systematically formulate and verify implicit visual claims against world knowledge.

2.3 Multimodal Misinformation Detection

Misinformation detection assesses the factual reliability of claims, articles, and social-media posts, and has evolved from text-only benchmarks (Wang, 2017; Thorne et al., 2018) to multimodal settings (Nakamura et al., 2020; Luo et al., 2021; Yao et al., 2023; Papadopoulos et al., 2024). Early methods capture event-invariant features or cross-modal consistency via end-to-end classifiers (Wang et al., 2018; Zhou et al., 2020; Nan et al., 2021). More recent systems introduce staged LLM/MLLM reasoning over internal knowledge or fixed corpora (Pan et al., 2023; Liu et al., 2025; Tahmasebi et al., 2024), while open-web frameworks retrieve multimodal evidence (Zhang & Gao, 2023; Qi et al., 2024; Xuan et al., 2024; Khaliq et al., 2024) and plan tool execution dynamically (Braun et al., 2025; Xu et al., 2026). Unlike these systems, which assess explicit textual claims or image–text consistency, K-VERA takes an uncaptioned image, formulates its implicit visual claims, and identifies AI generation through knowledge inconsistencies.

3 K-AIGI-Bench

To evaluate the ability of existing systems and our method to detect and explain knowledge inconsistencies, we construct K-AIGI-Bench. The benchmark contains $\sim$500 images generated by GPT-Image-2 and $\sim$500 topic-matched real images collected from the Internet. A diagram of the construction pipeline is provided in Appendix E.1 (Figure 6).

Knowledge topics. We define eight knowledge topics spanning stable facts, entity-specific attributes, general rules, and time-sensitive information. Figure 2 presents one AIGI from each topic: **Biology & Nature**, **Geography & Architecture**, **Products & Brands**, **History & Culture**, **Public Figures & Actions**, **Physics & Geometry**, **Software & Interfaces**, and **Events & Developments**. The same topics are used to organize both AIGIs and real images; they characterize the knowledge required for verification rather than an image's visual style or source generator. During generation, each AIGI is deliberately constructed with one knowledge inconsistency under its selected topic. Detailed definitions and additional examples are provided in Appendices E.2 and E.5, respectively.

AIGI construction. For each sample, we uniformly draw a knowledge topic and provide GPT-5.5 (OpenAI, 2026a) with summaries of 10 randomly sampled existing examples from that topic as references for diversity. GPT-5.5 constructs an image-generation prompt with one prespecified knowledge inconsistency and may use Web Search and Webpage Retrieval (Section 4.2) to obtain relevant, up-to-date information, especially for time-sensitive knowledge topics. Specifically, we require GPT-5.5 to generate and record the following information: (1) the image prompt and resolution, (2) a concise explanation of the knowledge inconsistency, (3) a short summary for diversity control, and (4) the required knowledge date, indicating the knowledge cutoff needed to identify the

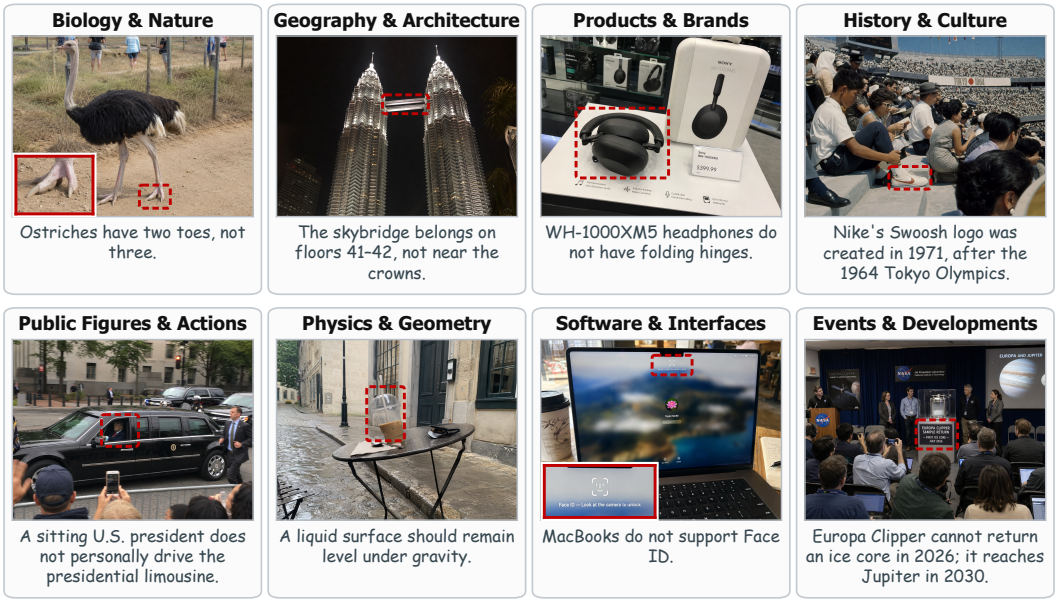

Figure 2: AIGI examples from the eight knowledge topics in K-AIGI-Bench. Each image contains one deliberately specified knowledge inconsistency, shown with a concise explanation.

inconsistency. We then use GPT-Image-2 to generate the image from the prompt and resolution. Further construction details are provided in Appendix E.1.

3.1 Topic-matched real images.

We collect ∼500 real images from the same eight knowledge topics. For example, Products & Brands includes recently released consumer products, whereas Geography & Architecture includes recognizable places and buildings. We retain only visually clear images with at least one spatial dimension exceeding 1,024 pixels and no watermark that reveals their source. We prioritize recent images from social-media platforms such as X[1] and Reddit[2]. For each real image, we estimate a required knowledge date and retain its source URL as metadata.

Human validation . All candidate images undergo manual validation before inclusion in K-AIGI-Bench. An AIGI is retained only if the prespecified knowledge inconsistency is visible, the image provides sufficient context to identify the implicit visual claim being tested, and no unrelated knowledge inconsistency or conspicuous semantic visual flaw is present. This procedure ensures that the recorded explanation is a meaningful target for determining whether a detector identified the intended evidence rather than a coincidental artifact. For each real image, reviewers examine its original publication context to confirm that it is a real image rather than an AI-generated one. Further validation details are provided in Appendix E.3.

4 K-VERA Approach

Detecting knowledge inconsistencies requires a system to move beyond image-intrinsic cues and verify implicit visual claims against external evidence. To this end, we introduce K-VERA (*Knowledge-grounded Visual Evidence Retrieval and Adjudication*), a training-free, retrieval-augmented multi-agent framework for knowledge-intensive AIGI detection. As illustrated in Figure 3, it follows a forensic workflow: resolve the image context, identify implicit visual claims, acquire evidence, and adjudicate the evidence.

[1] https://x.com/
[2] https://www.reddit.com/

4.1 TASK DEFINITION

Let I be an input image and $y \in \{0, 1\}$ its binary label, where $y = 1$ denotes fake and $y = 0$ denotes real. K-VERA considers the world knowledge available up to time T, including knowledge encoded in the backbone MLLM and knowledge acquired from external sources available no later than T. It returns $F_T(I) = (z, e, \mathcal{R})$, where z is a fine-grained verdict, e is a concise explanation, and $\mathcal{R} = \mathcal{R}^W \cup \mathcal{R}^I$ contains supporting webpages and reference images. We denote the four verdicts by $\mathcal{Z} = \{z_K, z_A, z_R, z_U\}$. Their definitions are as follows:

1. **Knowledge-Inconsistent** (z_K): The image contains no conspicuous semantic visual flaw, but at least one of its implicit visual claims conflicts with real-world knowledge. The conflict may be established using either the MLLM's internal knowledge or knowledge acquired through web retrieval.
2. **AIGI** (z_A): The image contains a conspicuous semantic visual flaw that provides direct visual evidence of AI generation.
3. **Real** (z_R): Based on web evidence, this image is most likely captured from reality rather than generated by AI.
4. **Unsure** (z_U): No knowledge inconsistency or conspicuous semantic visual flaw is established, and the available evidence is insufficient to support either a real or an AI-generated origin.

The set of fine-grained verdicts mapped to the fake class is $\mathcal{Z}^{\text{fake}} = \{z_K, z_A\}$, and the corresponding binary prediction is $\hat{y}_K = \mathbf{1}[z \in \mathcal{Z}^{\text{fake}}]$. Thus, an image is classified as fake when it contains either a conspicuous semantic visual flaw or a verified knowledge inconsistency. Although an "Unsure" verdict does not establish that the image is real, it is mapped to the real class in the binary evaluation.

4.2 RETRIEVAL TOOLS

K-VERA provides five tools: (1) *Web Search* retrieves webpages from text queries; (2) *Image Search* finds relevant images using text queries and returns their source-page URLs; (3) *Reverse Image Search* finds similar images using image queries and returns their source-page URLs; (4) *Webpage Retrieval* returns webpage content in Markdown format, preserving hyperlinks and image URLs; and (5) *Image Retrieval* displays images from URLs. Tool interfaces and implementation details are provided in Appendices F.1 and F.2. To prevent exact-match retrieval shortcuts, we use crop-resistant perceptual hashes to filter Image Search and Reverse Image Search results nearly identical to the input, while retaining related reference images.

4.3 AGENT DESIGN

K-VERA comprises four agent roles, as detailed below. Detailed prompts and input/output definitions are provided in Appendices A and F.3, respectively.

Context Resolver. Given the input image, it returns an uncertainty-aware description of the likely scene and identifiable entities, together with inspected webpages and reference images. It can use all five tools, prioritizing reverse image search for scene or entity identification, and treats retrieved context as a hypothesis rather than ground truth.

Planner. Given the image and the Context Resolver's description and references, it extracts implicit visual claims from the image and formulates a bounded set of independent verification tasks, accompanied by a concise reasoning summary. Each task contains an answerable question and the minimal context needed to investigate it, including relevant references. The Planner uses no tools.

Actors. Each Actor receives the image, one verification question, and its task-specific context. It independently returns a direct answer together with inspected webpages and reference images. Actors can use all five tools and favor primary or authoritative sources.

Adjudicator. Given the image, the Context Resolver's description and references, the Planner's reasoning summary, and all Actor answers and references, it checks source quality and support, resolves conflicts, and returns a fine-grained verdict, a concise explanation, and supporting webpages and reference images. It can use only Webpage Retrieval and Image Retrieval to inspect already cited sources.

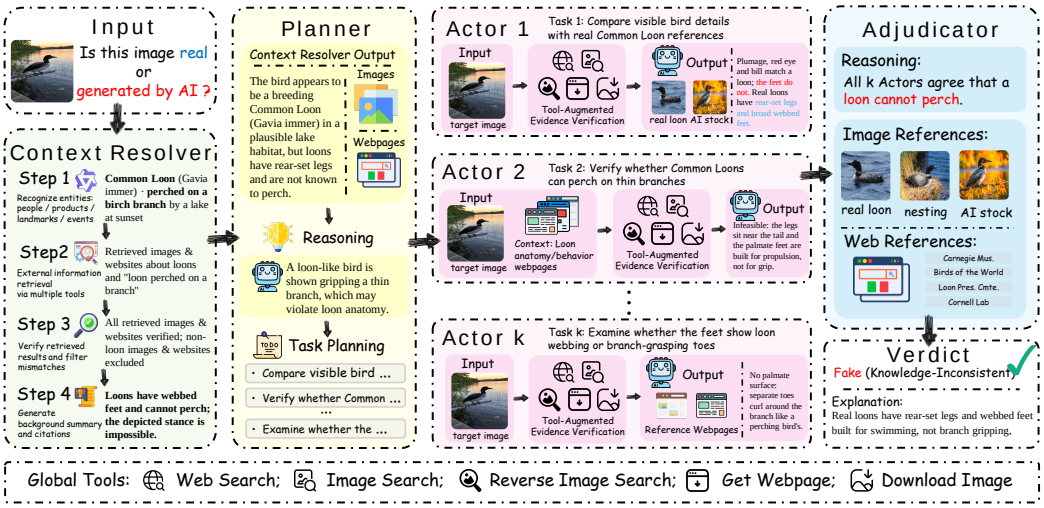

Figure 3: Overview of K-VERA. The Context Resolver identifies the scene and relevant entities, the Planner creates factual verification tasks, Actors acquire multimodal evidence in parallel, and the Adjudicator checks source support and returns a fine-grained verdict, explanation, and references.

4.4 Inference Workflow

K-VERA performs a single inference round. The Context Resolver establishes the image's likely context, and the Planner formulates independent verification questions about implicit visual claims. Actors then retrieve and assess multimodal evidence in parallel. Finally, the Adjudicator integrates their findings, checks source support, and resolves conflicts to produce a fine-grained verdict with an explanation and supporting reference webpages and images. The verdict is mapped to a binary real/fake prediction as defined in Section 4.1. A complete inference example is shown in Figures 12–16.

4.5 Complementing Pixel-Level Detectors

K-VERA complements pixel-level detectors through knowledge verification and auxiliary semantic visual cues, enabling their predictions to be combined without retraining. Given a conventional detector prediction $\hat{y}_D$, we use $\hat{y}_{\text{OR}} = \hat{y}_K \vee \hat{y}_D$. This rule preserves every fake prediction from either system. As demonstrated in Section 5, the resulting gain in fake accuracy outweighs the increase in the false-positive rate.

5 Experiments

5.1 Experimental Setup

Implementation details. We use Qwen3.7-Plus (Alibaba Cloud, 2026) as the default backbone MLLM with a decoding temperature of 0.01, and resize each input image under a maximum pixel budget of 1024×1024. Each tool-enabled agent may perform at most eight rounds of tool use, while the Planner may create at most four Actors. Web Search, Image Search, and Reverse Image Search each return up to eight results. All experiments use cutoff-controlled web retrieval with a search cutoff date of July 20, 2026.

Baselines. We compare K-VERA against three binary detectors—AIDE (Yan et al., 2025a), Effort (Yan et al., 2025b), and DDA (Chen et al., 2025a)—and three MLLM-based detectors—FakeVLM (Wen et al., 2025), FakeReasoning (Gao et al., 2025), and SIDA (Huang et al., 2025b). We additionally evaluate the Qwen3.7-Plus backbone without tools, prompting it to attend to both conspicuous semantic visual flaws and knowledge inconsistencies. We use the official pretrained

Table 1: Results on K-AIGI-Bench. Classification metrics and GT-KI Recall range from 0 to 1, while EQS ranges from 0 to 5 and is evaluated independently by Gemini 3.5 Flash, Kimi K2.6, and Qwen3.5-397B-A17B. Explanation metrics are not applicable to binary detectors.

Method	Real Acc.	Fake Acc.	BAcc.	F1	GT-KI Recall	EQS (0–5)		
						Gemini 3.5 Flash	Kimi K2.6	Qwen3.5-397B-A17B
AIDE	0.883	0.012	0.448	0.021	-	-	-	-
Effort	0.198	0.804	0.501	0.616	-	-	-	-
DDA	0.846	0.313	0.580	0.427	-	-	-	-
FakeVLM	0.530	0.361	0.445	0.394	0.002	0.40	0.40	0.40
FakeReasoning	**0.947**	0.034	0.490	0.062	0.002	0.02	0.02	0.02
SIDA	0.008	**1.000**	0.504	0.668	0.000	1.04	1.06	1.03
Qwen3.7-Plus	0.893	0.476	0.685	0.602	0.433	1.57	1.99	2.24
K-VERA	0.887	0.622	**0.754**	**0.716**	**0.562**	**2.17**	**2.71**	**2.82**

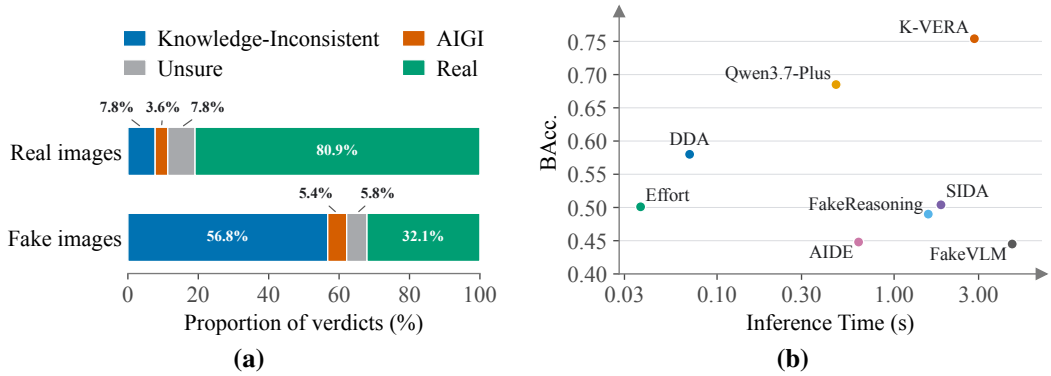

Figure 4: Analysis of K-VERA on K-AIGI-Bench. (a) Distribution of K-VERA's fine-grained verdicts on real and AI-generated images. (b) Balanced accuracy versus inference time across detectors. The horizontal axis in (b) is plotted on a logarithmic scale. All locally-run methods are evaluated on a single NVIDIA A100 40GB GPU, whereas Qwen3.7-Plus and K-VERA invoke APIs using 16 parallel processes.

models for all baselines; K-VERA and API-based baselines run on a CPU-only host, whereas locally deployed baselines run on one NVIDIA A100 40GB GPU.

Metrics. For classification, we report real accuracy, fake accuracy, balanced accuracy, and F1 with fake as the positive class. All classification metrics are reported on a 0–1 scale. To assess explanation quality, we further design two metrics for methods that produce explanations. Ground-Truth Knowledge-Inconsistency Recall (GT-KI Recall), also reported on a 0–1 scale, measures the fraction of AIGIs for which the generated explanation identifies the prespecified knowledge inconsistency, as evaluated by Qwen3.5-397B-A17B (Qwen Team, 2026). To account for knowledge inconsistencies or semantic visual flaws that may not have been identified during dataset construction, we introduce Explanation Quality Score (EQS) as a complementary, open-ended evaluation. EQS ranges from 0 to 5 and assesses whether an identified issue is factually correct, visibly grounded in the target image, and supported by valid evidence, with a higher score indicating a better explanation. We evaluate EQS independently with Gemini 3.5 Flash (Google DeepMind, 2026), Kimi K2.6 (Moonshot AI, 2026), and Qwen3.5-397B-A17B. All evaluators follow the same evaluation criteria and may use Webpage Retrieval and Image Retrieval to inspect provided evidence. The evaluation prompts are provided in Figures 26 and 27, and the detailed criteria in Appendix G.1.

5.2 Results on K-AIGI-Bench

Table 1 compares classification and explanation performance on the highly challenging K-AIGI-Bench. K-VERA achieves the highest balanced accuracy (0.754) and F1 score (0.716). Compared

Table 2: Ablation results on K-AIGI-Bench. Qwen3.7-Plus is used as the backbone MLLM.

Variant	BAcc.	F1	GT-KI Recall
Full K-VERA	**0.754**	**0.716**	**0.562**
w/o Context Resolver	0.738	0.692	0.543
w/o Web Search	0.727	0.667	0.519
w/o Image Search	0.746	0.689	0.537
w/o Reverse Image Search	0.749	0.713	0.544
w/o Image Retrieval	0.744	0.691	0.544

Table 3: Human evaluation of explanations. Samples with incorrect verdicts are automatically assigned a score of 0. FakeReasoning and SIDA are excluded because they show near-complete failure on K-AIGI-Bench.

Method	Score (0–5)
K-VERA	**4.76**
Qwen3.7-Plus	3.31
FakeVLM	0.39

with standalone Qwen3.7-Plus, K-VERA maintains comparable real accuracy (0.887 vs. 0.893) while improving fake accuracy from 0.476 to 0.622 and GT-KI Recall from 0.433 to 0.562, indicating that web search enables it to detect more knowledge inconsistencies. K-VERA also obtains the highest EQS under all three evaluators, which indicates that its explanations are more accurate and visibly grounded. By contrast, the MLLM-based baselines achieve low GT-KI Recall and EQS because they primarily search for conspicuous semantic visual flaws rather than verifiable knowledge inconsistencies, while such visual flaws are virtually absent from K-AIGI-Bench.

Figure 4(a) shows that K-VERA rarely predicts "AIGI" on K-AIGI-Bench, so its correct fake predictions arise almost entirely from "Knowledge-Inconsistent" verdicts. Figure 4(b) complements the classification results with the accuracy–inference-time trade-off. K-VERA requires more inference time than the other methods because it performs iterative retrieval and multi-agent verification, but achieves the highest balanced accuracy among them.

5.3 Combining MLLM-Based and Binary Detectors

We combine MLLM-based detectors with DDA using the OR rule in Section 4 and evaluate on three in-the-wild benchmarks: So-Fake-OOD (Huang et al., 2025c), GPT-Image-2 in the Wild (Zewde et al., 2026), and WildRF (Cavia et al., 2024). We randomly sample 500 real and 500 generated images from each of So-Fake-OOD and WildRF, reporting balanced accuracy, and 1,000 generated images from GPT-Image-2 in the Wild, reporting fake accuracy. We exclude So-Fake-OOD's tampered subset to focus on fully synthetic images. Further dataset details are provided in Appendix G.2. K-VERA OR DDA achieves the highest score on all three benchmarks (Table 4), consistently outperforming either component alone and supporting the complementarity of knowledge verification and pixel-level detection. SIDA's results are considered invalid because it classifies nearly all images as fake.

Table 4: Performance on three in-the-wild benchmarks. SIDA's results are shown in gray because it classifies almost every image as fake.

Method	So-Fake-OOD	GPT-Image-2	WildRF
AIDE	0.481	0.049	0.441
Effort	0.497	0.469	0.483
DDA	0.800	0.215	0.887
FakeVLM	0.668	0.409	0.708
FakeReasoning	0.489	0.141	0.531
SIDA	0.503	0.999	0.499
K-VERA	0.744	0.631	0.894
FakeVLM OR DDA	0.686	0.512	0.752
FakeReasoning OR DDA	0.776	0.316	0.877
K-VERA OR DDA	**0.846**	**0.707**	**0.907**

5.4 Ablation Study

We ablate the Context Resolver and four retrieval operations individually, with the results in Table 2 showing that each component contributes to the overall performance. Removing Web Search leads to the largest performance degradation across balanced accuracy, F1, and GT-KI Recall, confirming the importance of efficiently locating relevant textual evidence, while removing Reverse Image Search has the smallest effect, possibly because its functionality can be partially substituted by the other search operations. Removing the Context Resolver also degrades performance, as the Planner must begin planning without sufficient background information. Image Search and Image Retrieval are likewise beneficial, indicating that visual evidence provides complementary support beyond textual retrieval. See Appendix B for additional ablation experiments.

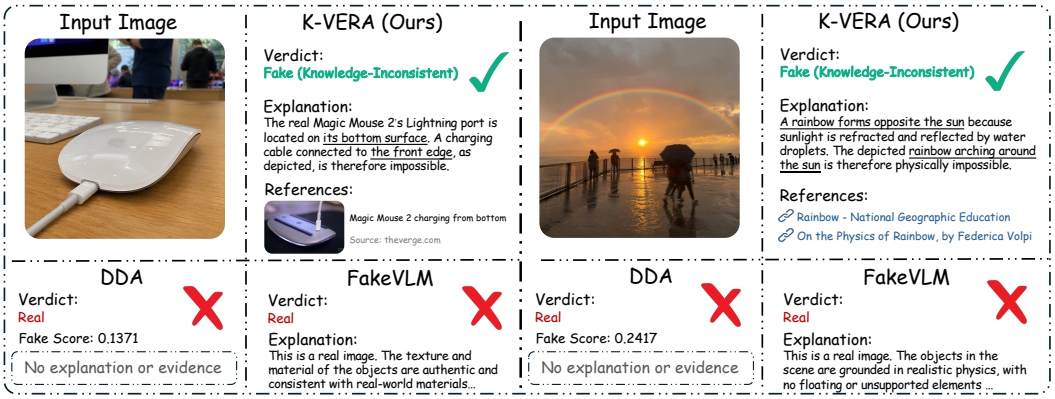

Figure 5: Qualitative comparison on two K-AIGI-Bench samples. K-VERA verifies their consistency against retrieved webpages and reference images, whereas DDA and FakeVLM incorrectly classify both images as real.

5.5 User Study

We conduct a human evaluation with 20 participants on explanations for images sampled from the fake subset of K-AIGI-Bench, with the results reported in Table 3. For each image, participants rate the explanations produced by K-VERA, Qwen3.7-Plus, and FakeVLM from 1 to 5: a score of 1 indicates that the explanation does not match the image or establish an AI-generated origin, whereas a score of 5 indicates a precise match and strong evidence of AI generation. A method's output receives a score of 0 when its binary verdict is incorrect. For a fair comparison, only explanation text is displayed, with K-VERA's reference webpages and images withheld. We exclude FakeReasoning and SIDA from the human evaluation because FakeReasoning almost always classifies images as real, whereas SIDA almost always classifies them as fake. The results demonstrate that K-VERA produces substantially higher-quality explanations from the perspective of human evaluators.

5.6 Qualitative Analysis

Figure 5 presents two representative examples. The binary detectors fail to classify the images as fake because the pixel-level characteristics of the new generator are unfamiliar, while the MLLM-based detectors find no semantic visual flaw and also return incorrect verdicts. K-VERA instead identifies a concrete implicit visual claim, retrieves authoritative webpages and reference images, and explains the visible conflict. These examples demonstrate K-VERA's substantial advantage over existing methods when modern-generator images contain knowledge inconsistencies but no conspicuous semantic visual flaws and exhibit previously unseen pixel-level characteristics.

6 Conclusion

We introduced knowledge-intensive AIGI detection, which identifies conflicts between implicit visual claims and real-world knowledge. This perspective complements image-intrinsic cues, including pixel-level artifacts and semantic visual flaws, especially for high-quality images from recent generators. K-AIGI-Bench provides a controlled evaluation spanning eight knowledge topics, with deliberately embedded knowledge inconsistencies for AIGIs, and topic-matched real images. K-VERA addresses the task through context resolution, focused verification, multimodal web retrieval, and evidence adjudication, without task-specific training. The experiments show that its evidence-grounded decisions improve both detection and explanation on K-AIGI-Bench and can strengthen a conventional detector through a simple OR rule. These results suggest that future AIGI detectors should evaluate not only how an image is rendered, but also whether what it depicts can be true.

AI USE STATEMENT

Generative AI tools were used as part of the research methodology to construct synthetic benchmark specifications, generate the AIGI subset of K-AIGI-Bench, execute the MLLM-based components of K-VERA, and evaluate explanation quality. Coding agents also assisted with browser-based collection of candidate public web images. In preparing the manuscript, generative AI tools assisted with literature search and summarization, drafting and language editing, reference formatting, and figure preparation. The authors checked literature claims and citations against the original sources, inspected retrieved evidence and model outputs, and verified the reported results from recorded experimental outputs. Generative AI was not used to formulate or prove mathematical theorems. The authors take responsibility for the final content of this work, including all text, claims, data, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is intended to improve the detection and explanation of AI-generated images that may otherwise be used as misleading visual evidence. Detection outputs remain fallible and should not be treated as definitive proof of authenticity or used as the sole basis for content moderation or decisions affecting individuals. K-AIGI-Bench contains deliberately fabricated visual claims and real images collected from publicly accessible web and social-media sources. We retain source URLs and provenance metadata and do not claim ownership of third-party content. The planned research release will document source-specific rights and usage limitations; third-party images will be provided in the form of source references or retrieval metadata. The benchmark and code will be released for research use, and removal requests concerning released examples will be honored. Although the benchmark includes public figures and public events, its examples are constructed for forensic evaluation rather than to target private individuals or promote the depicted false claims.

REPRODUCIBILITY STATEMENT

The construction and human-validation procedures for K-AIGI-Bench are described in Section 3; the task definition, retrieval tools, agent roles, and inference workflow are specified in Section 4; and the model settings, baselines, metrics, and evaluation protocols are reported in Section 5. Appendices E.1, E.3, F.1, and F.4 provide additional construction, validation, retrieval-tool implementation, and inference-workflow details, respectively. Upon publication, K-AIGI-Bench will be released for research use and the implementation of K-VERA will be released under the MIT License.

REFERENCES

Alibaba Cloud. Visual Understanding: Qwen3.7-Plus. https://www.alibabacloud.com/help/en/model-studio/vision-model, 2026. Accessed September 18, 2026.

Anthropic. Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, 2026. Accessed September 18, 2026.

Black Forest Labs. FLUX.1 [dev]. https://bfl.ai/blog/24-08-01-bfl, 2024. Accessed July 29, 2026.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. Defame: Dynamic evidence-based fact-checking with multimodal experts. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, 2025a.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. In *Advances in Neural Information Processing Systems*, 2025b.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Muxi Diao, Lei Chen, Kongming Liang, and Zhanyu Ma. Fakereasoning: Towards generalizable forgery detection and reasoning, 2025.

Google. Image editing in Gemini just got a major upgrade. `https://blog.google/products-and-platforms/products/gemini/updated-image-editing-model/`, 2025. Accessed July 29, 2026.

Google DeepMind. Gemini 3.5 Flash Model Card. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, 2026. Accessed September 18, 2026.

Qing Huang, Zhipei Xu, Xuanyu Zhang, Xiangyu Yu, and Jian Zhang. Unishield: An adaptive multi-agent framework for unified forgery image detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Tai-Ming Huang, Wei-Tung Lin, Kai-Lung Hua, Wen-Huang Cheng, Junichi Yamagishi, and Jun-Cheng Chen. Thinkfake: Reasoning in multimodal large language models for ai-generated image detection, 2025a.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, and Guangliang Cheng. So-fake: Benchmarking and explaining social media image forgery detection, 2025c.

Yikun Ji, Yan Hong, Qi Fan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In *International Conference on Learning Representations*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. Legion: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and Verification Workshop*, 2024.

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

Yixuan Li, Yu Tian, Yipo Huang, Wei Lu, Shiqi Wang, Weisi Lin, and Anderson Rocha. Fakescope: Large multimodal expert model for transparent ai-generated image forensics, 2025b.

Mengfei Liang, Yiting Qu, Yukun Jiang, Michael Backes, and Yang Zhang. From evidence to verdict: An agent-based forensic framework for ai-generated image detection, 2025.

Xuannan Liu, Zekun Li, Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *International Conference on Learning Representations*, 2025.

Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Moonshot AI. Kimi K2.6. `https://www.kimi.ai/ai-models/kimi-k2-6`, 2026. Accessed September 18, 2026.

Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 2021.

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

OpenAI. GPT-5.5. `https://developers.openai.com/api/docs/models/gpt-5.5`, 2026a. Accessed September 18, 2026.

OpenAI. GPT-5.6 Sol. `https://developers.openai.com/api/docs/models/gpt-5.6-sol`, 2026b. Accessed September 18, 2026.

OpenAI. GPT Image 2. `https://developers.openai.com/api/docs/models/gpt-image-2`, 2026c. Accessed July 29, 2026.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, Panagiotis C. Petrantonakis, et al. Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(4), 2024.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Qwen Team. Qwen3.5-397B-A17B. `https://huggingface.co/Qwen/Qwen3.5-397B-A17B`, 2026. Accessed September 18, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems*, 2025.

Danni Xu, Shaojing Fan, Harry Cheng, and Mohan Kankanhalli. Rw-post: Auditable evidence-grounded multimodal fact-checking in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, 2025.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation, 2024.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, 2025a.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

Yangxin Yu, Yue Zhou, Bin Li, Kaiqing Lin, Haodong Li, Jiangqun Ni, and Bo Cao. Agentfox: Llm agent-guided fusion with explainability for ai-generated image detection, 2026.

Kidus Zewde, Simiao Ren, Xingyu Shen, Jiaqi Wu, Yuchen Zhou, Tommy Duong, Zikang Zhang, Ethan Traister, and Kewen Xie. Gpt-image-2 in the wild: A twitter dataset of self-reported ai-generated images from the first week of deployment, 2026.

Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2023.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2020.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Chenyang Zhu, Maorong Wang, Jun Liu, Ching-Chun Chang, and Isao Echizen. Evoguard: An extensible agentic rl-based framework for practical and evolving ai-generated image detection, 2026.

APPENDIX

A PROMPTS

We summarize the prompt templates used in this work below.

1. **Benchmark construction:** the prompt for generating image specifications and metadata with GPT-5.5 (Figures 17–18).
2. **Context Resolver:** system and user prompts for identifying the image context (Figure 19).
3. **Planner:** system and user prompts for extracting implicit visual claims and planning verification tasks (Figures 20–21).
4. **Actor:** system and user prompts for investigating individual verification tasks (Figure 22).
5. **Adjudicator:** system and user prompts for integrating evidence and producing the final verdict and explanation (Figures 23–24).
6. **Tool-free backbone:** system and user prompts for the backbone baseline without retrieval tools (Figure 25).
7. **GT-KI Recall evaluation:** system and user prompts for matching predicted explanations to ground-truth knowledge inconsistencies (Figure 26).
8. **EQS evaluation:** system and user prompts for assessing explanation quality (Figure 27).

B ADDITIONAL ABLATION STUDIES

We vary the maximum number of tool-use rounds per tool-enabled agent and the maximum number of Actors, holding the other limit at its default value. Table 5 summarizes the results.

Table 5: Sensitivity to inference budgets.

Max. rounds	Max. Actors	Real Acc.	Fake Acc.	BAcc.	F1
4	4	0.859	0.647	0.753	**0.723**
8	4	**0.887**	0.622	**0.754**	0.716
16	4	0.837	**0.657**	0.747	0.722
8	1	0.869	0.596	0.732	0.690
8	2	0.808	0.653	0.731	0.707
8	8	0.838	0.643	0.741	0.712

The additional runs achieve balanced accuracies between 0.731 and 0.753, including configurations with fewer tool-use rounds or Actors. These results suggest that knowledge-based verification remains useful under smaller inference budgets, supporting its use alongside conventional detectors when computational resources are limited.

C ROBUSTNESS TO IMAGE DEGRADATION

We examine JPEG compression at quality 70, Gaussian blur with $\sigma = 1$, and resizing each spatial dimension by a factor of 0.5. We also evaluate the resized images after enlarging them back to their original dimensions. Table 6 compares K-VERA with the three binary detectors.

Table 6: Balanced accuracy under image degradation.

Condition	K-VERA	AIDE	Effort	DDA
Clean	**0.754**	0.448	0.501	0.580
JPEG (quality 70)	**0.758**	0.468	0.446	0.494
Gaussian blur ($\sigma = 1$)	**0.758**	0.450	0.362	0.474
Downsampling ($\times 0.5$)	**0.717**	0.460	0.459	0.471
Downsampling then upsampling	**0.747**	0.443	0.336	0.547

JPEG compression and Gaussian blur have little effect on K-VERA's performance, whereas downsampling reduces its balanced accuracy to 0.717. However, enlarging the downsampled images back to their original dimensions restores much of the performance, raising balanced accuracy to 0.747. Since upsampling does not restore lost image information, this recovery suggests that input resolution and the associated allocation of visual tokens may influence fine-detail recognition. Preserving sufficient input resolution can therefore help knowledge verification even after image degradation.

D BACKBONE MLLMS

K-VERA does not update its backbone parameters and can use different MLLMs without changing the workflow. Table 7 compares Qwen3.7-Plus with three leading proprietary models—GPT-5.6-Sol (OpenAI, 2026b), Gemini 3.5 Flash (Google DeepMind, 2026), and Claude Opus 4.8 (Anthropic, 2026)—on K-AIGI-Bench. All three alternative backbones improve balanced accuracy over Qwen3.7-Plus, indicating that K-VERA can achieve further performance gains when equipped with stronger backbone models.

Table 7: Classification performance of K-VERA with different backbone MLLMs on K-AIGI-Bench.

Backbone	Real Acc.	Fake Acc.	BAcc.	F1
Qwen3.7-Plus	0.887	0.622	0.754	0.716
GPT-5.6-Sol	**0.949**	0.657	0.803	0.769
Gemini 3.5 Flash	0.788	**0.826**	**0.807**	**0.811**
Claude Opus 4.8	0.929	0.667	0.798	0.767

E K-AIGI-BENCH DETAILS

E.1 CONSTRUCTION PIPELINE

Figure 6 summarizes the construction process. GPT-5.5 receives a randomly selected knowledge topic and summaries of up to ten existing examples from that topic; fewer examples are supplied when the topic archive has not yet accumulated ten entries. It constructs a new image specification, optionally checking background or time-sensitive facts with Web Search and Webpage Retrieval. GPT-Image-2 then renders the specification, and human validation determines whether the sample is retained. The construction prompt is provided in Figures 17–18.

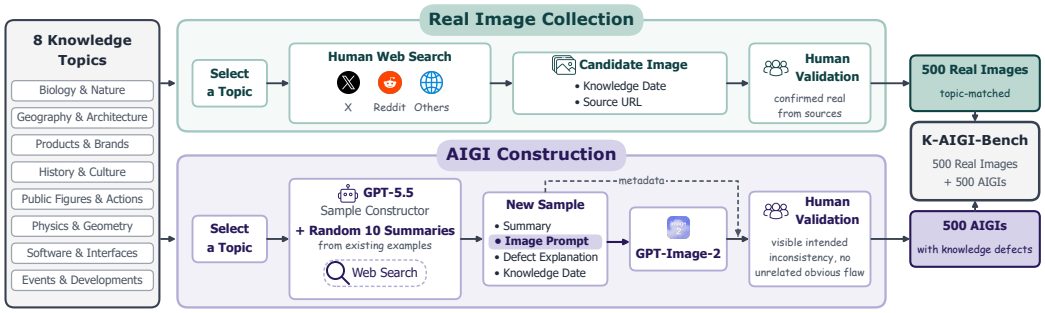

Figure 6: Construction pipeline of K-AIGI-Bench. Topic selection, reference lookup, and diversity control guide prompt construction. GPT-Image-2 renders candidate AIGIs, which are manually validated with their metadata. Topic-matched real images are collected and verified separately.

The constructor chooses a presentation format, such as an on-site photograph, news screenshot, social-media post, or product advertisement, and explicitly describes it in the image prompt. Width and height are multiples of 64 between 1,024 and 2,048 pixels. The prompt asks for a plausible image without self-revealing labels such as "AI generated" and without unintended flaws. Search is useful, for example, to verify that a supposedly fabricated event has not actually occurred or that a supposedly nonexistent product has not been released.

The recorded specification includes the image prompt, resolution, defect explanation, presentation format, diversity-control summary, and required knowledge date. The required knowledge date

specifies how recent the knowledge must be to identify the inconsistency: the earliest relevant year–month for dated facts, "before 2000" for stable facts, or "now" for facts that may change. Images are converted to JPEG with quality 90 to reduce file-format and compression differences as label shortcuts.

E.2 Knowledge-Topic Definitions

The eight topics describe the knowledge associated with image content and apply to both real and generated images.

1. **Biology & Nature:** organisms and their morphology, habitats, behavior, and growth, together with other properties of the natural world.
2. **Geography & Architecture:** geographic locations, landmarks, buildings, and their appearances and spatial layouts.
3. **Products & Brands:** identifiable products and brands, including their components, ports, logos, designs, and technical specifications.
4. **History & Culture:** historical periods and cultural settings, including their clothing, objects, technologies, architecture, and practices.
5. **Public Figures & Actions:** public figures and facts about their identities, offices, status, activities, and interactions.
6. **Physics & Geometry:** physical and structural properties, including gravity, reflections, shadows, spatial geometry, mechanics, and the rules of structured objects such as puzzles.
7. **Software & Interfaces:** operating systems, software capabilities, interface layouts and states, and interaction logic.
8. **Events & Developments:** events, product releases, crises, discoveries, and other developments, including their occurrence, timing, and factual context. Product attributes are covered by Products & Brands, whereas a product's release is covered by this topic.

E.3 Real Images and Human Validation

Real images are collected from the same topics, prioritizing recent posts on X and Reddit. Candidates must be visually clear, have at least one spatial dimension exceeding 1,024 pixels, and lack a watermark revealing their source. Their source URLs and estimated required knowledge dates are retained. Reviewers inspect the original publication context to assess whether the image is a real photograph rather than AI-generated content.

For each AIGI, reviewers check that the intended inconsistency is visible, that the image provides enough context to identify the claim, and that no unrelated knowledge inconsistency or conspicuous semantic visual flaw is apparent. The explanation must describe the retained image rather than merely the intended prompt. Candidates that do not meet these criteria are excluded.

E.4 Dataset Statistics

In total, K-AIGI-Bench contains 1,010 images: 506 topic-matched real images and 504 images generated by GPT-Image-2. Figures 7 and 8 show the distributions of knowledge topics and required knowledge dates, reported separately for real and generated images. Counts are grouped by knowledge topic and required knowledge date, respectively.

E.5 Additional Examples

Figure 9 presents additional examples across all eight topics. Each panel pairs the image with its topic and ground-truth inconsistency explanation.

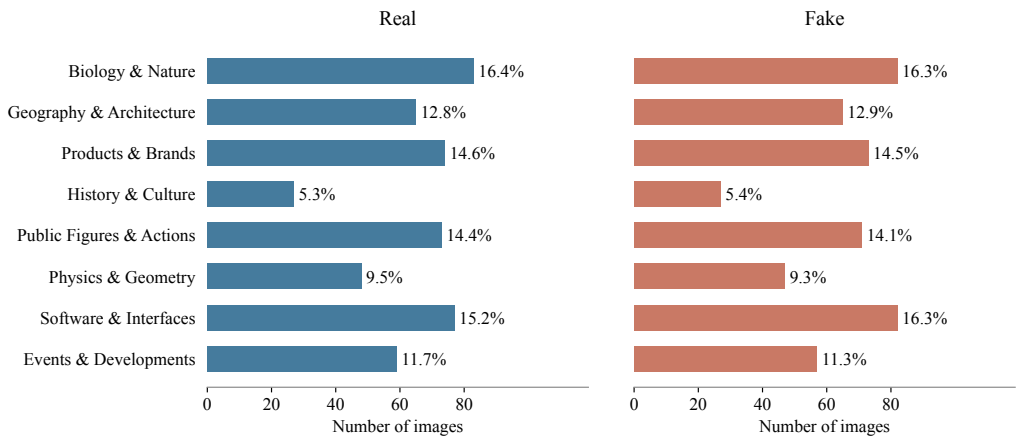

Figure 7: Knowledge-topic distributions in K-AIGI-Bench.

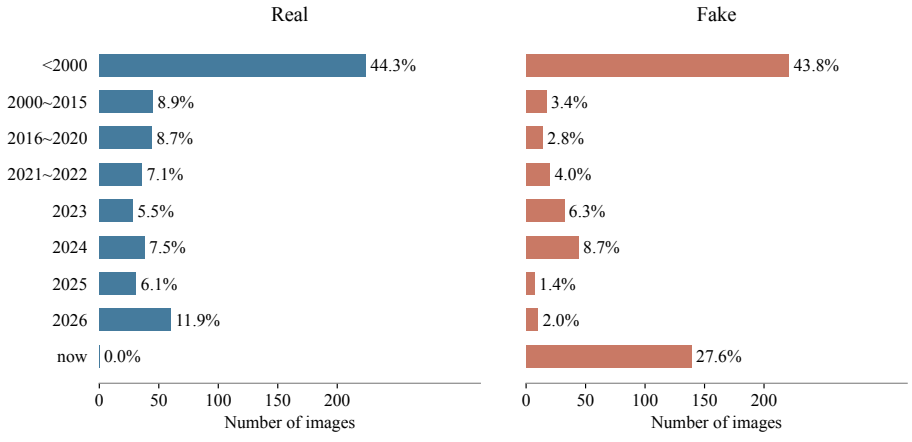

Figure 8: Required knowledge date distributions. The "now" category denotes facts that have not yet occurred but may change at any time in the future.

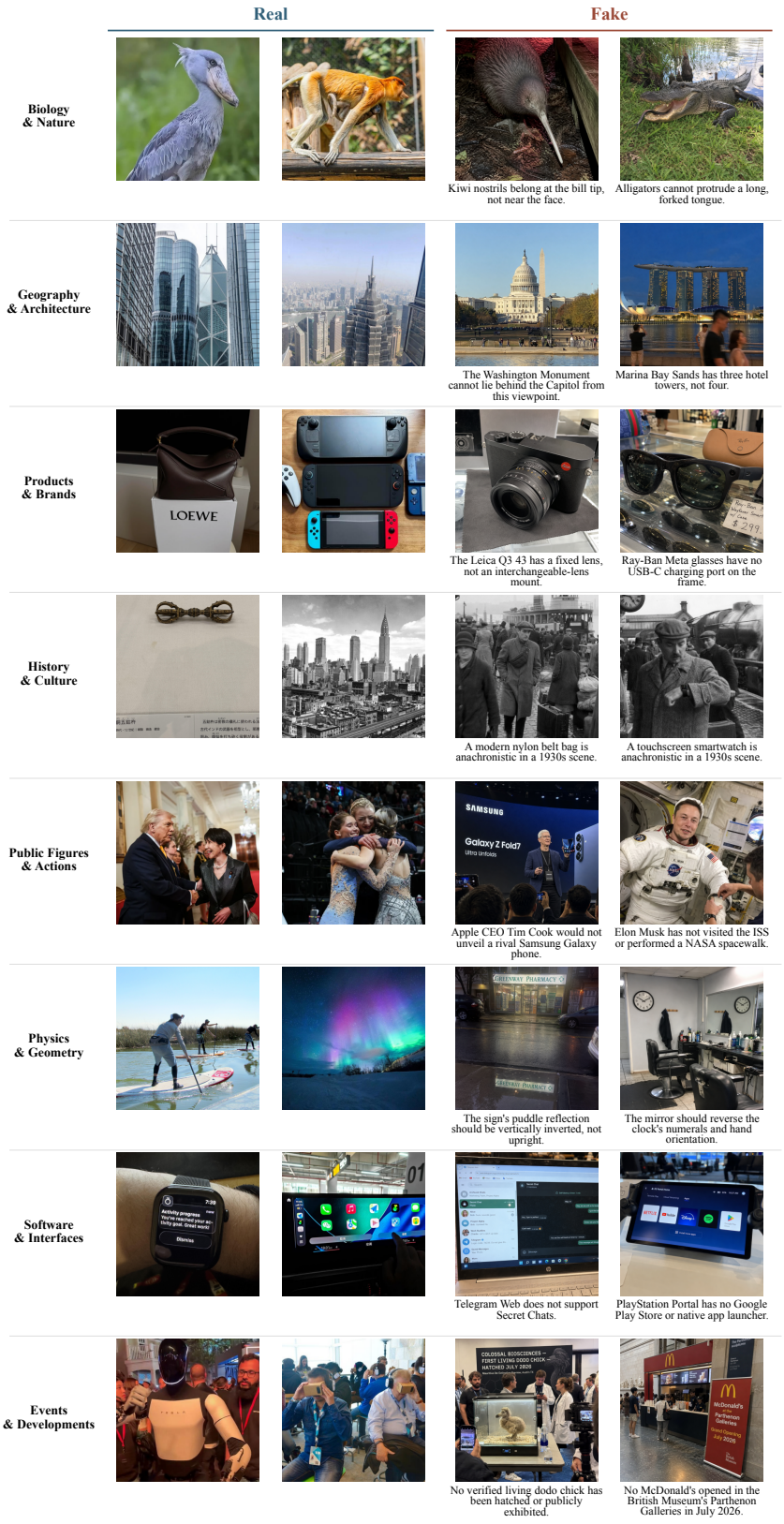

Figure 9: Additional K-AIGI-Bench examples.

F K-VERA IMPLEMENTATION DETAILS

F.1 TOOL INTERFACES AND SERVICES

The five tools provide the following inputs and outputs:

1. **Web Search** accepts a text query and returns ranked results in Markdown, including titles, webpage URLs, snippets, and source or date information when available.
2. **Image Search** accepts a text query and returns image titles, direct image URLs, source-page URLs, and a composite preview.
3. **Reverse Image Search** accepts an image and an optional region of interest. It searches with the whole image or the selected crop and returns image results and a composite preview in the same format as Image Search.
4. **Webpage Retrieval** accepts a webpage URL and an optional range of content blocks. It returns readable Markdown with hyperlinks and image URLs preserved. Pages are divided into 2,000-character blocks so that agents can selectively inspect their content.
5. **Image Retrieval** accepts an image URL and returns the downloaded image for visual inspection.

Every tool response appends the number of remaining tool-call rounds, reminding the agent how many rounds it may still use within its current stage.

Web Search, Image Search, and Reverse Image Search use the Serper API[3]. Reverse search requires a public image URL, so query images or crops are hosted using Cloudinary[4]. Webpage Retrieval uses Crawl4AI[5]. Search tools return up to eight results by default, and image-search previews use a 4×2 grid. The search service applies a date-range filter using the configured cutoff date.

F.2 NEAR-DUPLICATE IMAGE FILTERING

We filter near-duplicate search results using crop-resistant perceptual hashes from the `ImageHash` library.[6] The bit-error-rate threshold is 0.1, and a candidate is removed when at least one region matches the input in either comparison direction.

F.3 AGENT INPUTS, OUTPUTS, AND TOOL PERMISSIONS

Let I denote the image and $\mathcal{R}^I$ and $\mathcal{R}^W$ denote sets of reference images and webpages. A webpage reference contains a title and source URL; an image reference additionally contains a direct image URL. The corresponding prompt templates appear in Figures 19–24.

Context Resolver. The mapping $C(I) = c = (s_c, \mathcal{R}^I_c, \mathcal{R}^W_c)$ returns an uncertainty-aware scene/entity description s_c and inspected references. All five tools are available. Reverse image search is prioritized when it can resolve the scene or an entity; retrieved information remains a hypothesis to verify.

Planner. The mapping $P(I, c) = (s_p, \mathcal{T})$ returns a reasoning summary s_p and tasks $\mathcal{T} = \{t_j = (q_j, c_j)\}_{j=1}^m$ with $m \leq N_A$. Each question q_j concerns an implicit visual claim extracted from I; its optional context c_j contains only relevant background and references. The Planner has no tools.

Actors. An Actor maps (I, q_j, c_j) to $r_j = (s_j, \mathcal{R}^I_j, \mathcal{R}^W_j)$, where s_j directly answers the question and the reference sets contain inspected evidence. Actors operate independently with access to all five tools.

Adjudicator. The mapping $A(I, c, s_p, \{r_j\}_{j=1}^m) = (z, e, \mathcal{R})$ returns the fine-grained verdict, concise explanation, and supporting references. It checks source support and conflicts and briefly examines the image for conspicuous semantic visual flaws. Only Webpage Retrieval and Image Retrieval are available, for inspecting already cited sources. The four verdicts and binary mapping follow Section 4.1.

[3] `https://serper.dev/`
[4] `https://cloudinary.com/`
[5] `https://github.com/unclecode/crawl4ai`
[6] `https://github.com/JohannesBuchner/imagehash`

F.4 Inference Workflow

K-VERA performs a single inference round:

1. **Context resolution.** The Context Resolver inspects the input image I and uses retrieval tools to establish its likely scene, location, time, and entities. It returns $c = C(I) = (s_c, \mathcal{R}_c^I, \mathcal{R}_c^W)$, comprising an uncertainty-aware description and inspected reference images and webpages.
2. **Task planning.** Given (I, c), the Planner extracts implicit visual claims and decomposes them into independent verification questions. It returns $(s_p, \mathcal{T}) = P(I, c)$, where s_p summarizes its reasoning and $\mathcal{T} = \{(q_j, c_j)\}_{j=1}^m$ contains at most N_A tasks, each with a question and relevant context and references.
3. **Evidence acquisition.** Parallel Actors independently investigate the tasks in $\mathcal{T}$. Each Actor uses the available tools to consult reliable sources and compare the image with relevant references, mapping (I, q_j, c_j) to $r_j = (s_j, \mathcal{R}_j^I, \mathcal{R}_j^W)$, a direct answer with inspected images and webpages.
4. **Evidence adjudication.** The Adjudicator integrates the image, context, Planner reasoning, and Actor findings to compute $(z, e, \mathcal{R}) = A(I, c, s_p, \{r_j\}_{j=1}^m)$. It inspects cited sources when needed, checks evidence support, resolves conflicts, and briefly checks the image for conspicuous semantic visual flaws. The final output is $F_T(I) = (z, e, \mathcal{R})$, with references supporting the explanation; the verdict is mapped to a binary prediction as defined in Section 4.1.

Each tool-enabled role may make multiple tool-use rounds within its stage. Figures 12–16 present a complete inference example, tracing the same input image and references across all four stages to connect the final verdict to the verification tasks and inspected evidence. The evidence-acquisition stage is shown separately for the two parallel Actors. System prompts are omitted because they are already provided in Appendix A. To save space, the displayed text is also condensed.

G Evaluation Details

The backbone, inference settings, baselines, and principal metrics are described in Section 5.1. We provide the explanation-evaluation criteria here and the corresponding prompts in Figures 26 and 27.

G.1 Explanation Evaluation

GT-KI Recall. The evaluator receives the ground-truth defect explanation and the predicted explanation. It returns a Boolean semantic match and a brief reason. A match requires identifying the same concrete defect or factual inconsistency; a generic claim that an image is fake is insufficient. Different wording and additional observations are allowed. Recall is the fraction of evaluated fake-image explanations that match their ground truth.

EQS. Fake images assigned an incorrect binary verdict receive zero without a quality-evaluator call. For the remaining outputs, the evaluator receives the image, explanation, and supplied webpage/image references and assigns an integer from 1 to 5 according to Table 8. Scores are averaged separately for each evaluator, including the zero scores. A cited URL is not itself evidence that the associated claim is correct: the visible claim and the supporting source must be checked.

G.2 External Benchmarks

So-Fake-OOD (Huang et al., 2025c) contains real, fully synthetic, and tampered images collected from Reddit or generated by commercial models. We exclude the tampered subset to focus on fully synthetic images and randomly sample 500 real and 500 generated images.

GPT-Image-2 in the Wild (Zewde et al., 2026) contains 10,217 self-reported GPT-Image-2 images collected from X shortly after the generator's release. We randomly sample 1,000 generated images and report fake accuracy because this dataset does not provide a corresponding real-image class for our evaluation.

WildRF (Cavia et al., 2024) contains real and generated images from Reddit, X, and Facebook, reflecting platform processing and unknown generation pipelines. We randomly sample 500 images per class and report balanced accuracy, as for So-Fake-OOD.

Table 8: EQS scoring criteria. Scores 1–5 are assigned by the evaluator; score 0 is assigned by the evaluation procedure.

Score	Criterion
0	Incorrect binary verdict on a fake image.
1	Claimed defects are unclear or absent, or the evidence is nonexistent, irrelevant, or wholly unreliable.
2	Limited support: the main observations or evidence are weak or ambiguous.
3	Meaningful support with notable uncertainty, weak evidence, or partially unsupported claims.
4	Substantially convincing, with only minor omissions or weaknesses.
5	Clear, important defects or inconsistencies supported by strong, valid evidence.

H Discussion on the Unsure Verdict

We introduced the "Unsure" verdict option for two main reasons. First, the diversity of the real world and the barriers of web search—some websites do not allow direct access by unregistered users—make it inevitable that some images cannot have their authenticity confirmed directly through online evidence; for such images, returning "Unsure" is reasonable and accurate. Second, in our experiments we observed that the more an MLLM deliberates, the more prone it becomes to overthinking, often questioning subtle details in real images and thus misclassifying them as fake; offering the "Unsure" option alleviates such overthinking. Mapping "Unsure" to real is also consistent with a principle analogous to the presumption of innocence: if no defect of an image can be confirmed, the image should be regarded as real. Figure 10 shows two examples in which K-VERA returns "Unsure" for real images and is therefore correct under the binary evaluation.

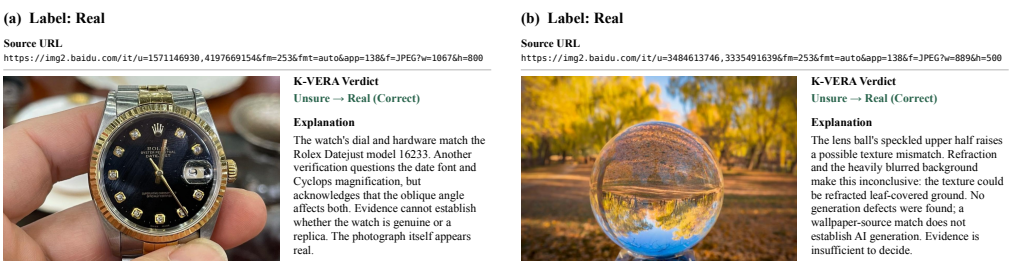

Figure 10: Two examples in which K-VERA returns "Unsure" for real images and is therefore correct under the binary evaluation, which maps "Unsure" to the real class.

I Reliability of Retrieved Evidence

Our prompts instruct agents to prefer primary or authoritative sources, distinguish confirmed facts from uncertainty, and check whether cited evidence supports the conclusion. K-VERA does not add a separate source-trustworthiness classifier or an independent reliability-verification service. It relies in part on the search engine's ranking and information-quality mechanisms: Google describes using relevance and quality signals, including signals associated with expertise and trustworthiness, to prioritize reliable information.[7] Search-engine ranking and agent-level evidence assessment jointly support source selection. Dedicated source verification would offer a further route to improving reliability at additional retrieval and inference cost.

[7] https://blog.google/products-and-platforms/products/search/how-google-delivers-reliable-information-search/

J Relationship Between Different AIGI Detection Signals

We compare the binary predictions of FakeVLM, FakeReasoning, and K-VERA with those of DDA, computed separately within each dataset and ground-truth class. Fake is encoded as 1 and real as 0 following the binary mapping in Section 4.1. Pearson correlation between these binary variables equals the phi coefficient.

Table 9: Phi coefficient with DDA's binary predictions, computed on paired images within each dataset and true class.

Detector	So-Fake-OOD		GPT-Image-2 in the Wild	WildRF	
	Real	Fake	Fake	Real	Fake
FakeVLM	0.051	0.106	0.119	0.008	0.146
FakeReasoning	0.017	−0.036	0.068	0.082	0.035
K-VERA	0.060	0.023	0.017	0.069	0.038

The coefficients are all close to zero (at most 0.146 in magnitude), indicating that the MLLM-based methods rely on signals essentially independent of the pixel-level signal captured by DDA. As a result, an MLLM-based method can serve as a complementary detection signal for fake images, provided that its false-positive rate on real images is low: combined with a pixel-level detector, it can recover fake images that the pixel-level detector misses without substantially increasing false positives. The combination results in Section 5.3 quantify these gains.

K Failure Cases and Limitations

Figure 11 presents four failure cases with the target image, ground truth, and system output. Model hallucinations, incorrect entity identification, and unreliable or misinterpreted retrieval results can lead to incorrect verdicts.

Knowledge-based detection also has an intrinsic limitation: a real photograph can depict an unusual scene that appears inconsistent without its publication context. For example, a film-set photograph may show historical clothing and buildings alongside modern electronic equipment. This combination does not establish AI generation. Context identifying the location as a film set can resolve the apparent contradiction and reduce false positives.

Conversely, an AI-generated image need not contain any detectable knowledge inconsistency. Knowledge verification therefore complements other detection signals rather than covering all AIGIs on its own. Retrieval and multi-agent inference also incur greater latency and monetary cost than lightweight binary detection, and their effectiveness depends on the availability and quality of external evidence.

(a) Label: Real

Source URL
https://www.reddit.com/r/DOG/comments/1sbphx1/my_rescue_dog_got_his_very_own_photoshoot_last/



K-VERA Verdict
AIGI (Incorrect)

Explanation
Daffodils allegedly have more than six tepals and malformed petals. Even frontal lighting despite strong backlighting is considered suspicious. No reverse-image match establishes provenance. Together these are taken as strong evidence of AI generation.

Reference Images (None)

Reference Webpages
1. Narcissus (plant) - Wikipedia
   https://en.wikipedia.org/wiki/Narcissus_(plant)
2. Narcissus - Britannica
   https://www.britannica.com/plant/narcissus-plant

(b) Label: Real

Source URL
https://img0.baidu.com/it/u=2999427277,971663321&fm=253&fmt=auto&app=138&f=JPEG?w=750&h=500



K-VERA Verdict
Knowledge-Inconsistent (Incorrect)

Explanation
The lens inscription matches the Fujifilm X100V, but the controls are claimed to be on the LEFT rather than the RIGHT as in reference photos. Correctly readable lens text is taken to rule out mirroring, implying an impossible control layout.

Reference Images


Fujifilm X100V silver top plate (Macfilos)


Fujifilm X100V silver top view (Ken Rockwell)

Reference Webpages
1. Fujifilm X100V Parts Diagram
   https://fujifilm-dsc.com/en/manual/x100v/about_this_camera/parts/index.html

(c) Label: Fake

Ground-Truth Defect
Loggerhead sea turtle hatchlings do not have five separate clawed toes on each front flipper; their foreflippers are paddle-like and bear only one or two external claws.



K-VERA Verdict
Real (Incorrect)

Explanation
Identified as a snapping turtle hatchling, not a sea turtle: its keeled shell, clawed digits and hooked beak are deemed normal. A sandy shore with waves is explained as a plausible Great Lakes habitat. No visual flaws or evidence of AI generation are found.

Reference Images


Snapping turtle hatchling on sand (Todd Henson Photography)


Baby snapping turtle on Lake Superior shore (Reddit)

Reference Webpages
1. Snapping Turtle Hatchling — Todd Henson Photography
   https://toddhensonphotography.com/blog/snapping-turtle-hatchling
2. Baby snapping turtle on Lake Superior shore in L'Anse, Michigan (Reddit)
   https://www.reddit.com/r/Michigan/comments/htpi2x/baby_snapping_turtle_on_lake_superior_shore_in/
3. Snapping Turtle (Chelydra serpentina) — Montana Field Guide
   https://fieldguide.mt.gov/speciesDetail.aspx?elcode=araab01010

(d) Label: Fake

Ground-Truth Defect
A real Raspberry Pi 5 has two micro-HDMI display ports, not a single full-size HDMI Type-A connector.



K-VERA Verdict
Real (Incorrect)

Explanation
Judged a genuine Raspberry Pi 5 matching official specifications and retail photos. The model claims dual micro-HDMI ports, with the second obscured by the cable and viewing angle. USB ports, GPIO pins and component placement are considered correct.

Reference Images


Raspberry Pi 5 4GB eBay listing photo


Raspberry Pi 5 Amazon product photo

Reference Webpages
1. Raspberry Pi 5 official product page
   https://www.raspberrypi.com/products/raspberry-pi-5/
2. Introducing Raspberry Pi 5 — official specs
   https://www.raspberrypi.com/news/introducing-raspberry-pi-5/

Figure 11: Failure cases of K-VERA.

Context Resolver

01 User prompt

Resolve only the background context needed to assess the attached target image. Keep the result concise and do not decide whether the image is real or fake.



image_0

02 Assistant thinking

... This appears to be Mount Rushmore, but I initially count FIVE faces. The standard monument has four. Recounting is uncertain; use reverse image search and check the landmark's appearance. ...

03 Tool call

```
reverse_image_search({"image_label": 0})
search({"query": "Mount Rushmore how many presidents faces carved"})
```

04 Tool response

Reverse image search: 8 results.
1. Wikipedia: File:MtRushmore.jpg [W4], image [I1].
... Other Mount Rushmore photos from travel and social sites. ...
Preview: image_1.

05 Tool response

Web search: 8 results.
1. NPS, 'Why These Four Presidents?' [W1]: Borglum selected four presidents.
... 6. History.com [W2]: four gigantic presidential sculptures. ...
Tool rounds remaining: 7.

06 User / tool images

Images returned by the preceding tool calls.



image_1

07 Assistant thinking

... Search confirms Mount Rushmore. I now see four faces and consider the layout consistent. The retrieved photos show similar viewpoints, not the exact target image. Enough background context; no further search. ...

08 Output

```
{
  "context_information": {
    "text": "The image depicts Mount Rushmore National
    Memorial in the Black Hills of South Dakota, USA. ...
    The four faces shown match the actual monument's layout
    (Washington on the far left, then Jefferson, Roosevelt,
    and Lincoln on the far right).",
    "reference_images": [{"title": "...", "image_url":
    [I1], "link": [W4]}],
    "reference_webpages": [{"title": "...", "link": [W1]},
    {"title": "...", "link": [W2]}]
  }
}
```

```
[I1] https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/MtRushmore.jpg/1280px-MtRushmore.jpg
[W1] https://www.nps.gov/moru/learn/historyculture/why-these-four-presidents.htm
[W2] https://www.history.com/articles/mount-rushmore
[W4] https://en.wikipedia.org/wiki/File:MtRushmore.jpg
```

Figure 12: Complete example, stage 1: context resolution.

Planner

01 User prompt

Target-image context:


```
{
 "text": "The image depicts Mount Rushmore National Memorial in
  the Black Hills of South Dakota, USA. ... The four faces shown
  match the actual monument's layout (Washington on the far left,
  then Jefferson, Roosevelt, and Lincoln on the far right).",
 "reference_images": [{"title": "...", "image_url": [I1], "link": [W4]}],
 "reference_webpages": [{"title": "...", "link": [W1]}, {"title": "...",
  "link": [W2]}]
}
```


Extract implicit visual claims from the attached target image and plan their verification. Return no more than 4 independent tasks. Include relevant references in each task's minimal context.



image_0

02 Assistant thinking

... A quick count suggests five faces rather than four, but I repeatedly reconsider the identities. Verify the count and arrange a comprehensive reference-photo comparison of order, mustaches, beards, relative sizes and repeated/deformed appearances. ...

03 Output


```
{
  "planner_reasoning": "The target image appears to show
  Mount Rushmore, but a quick visual count suggests there
  are five sculpted faces instead of the actual four. ...",
  "verification_tasks": [{
    "question": "Count the number of distinct sculpted
    presidential faces visible on the mountain in the
    target image ... exactly four faces (Washington,
    Jefferson, T. Roosevelt, and Lincoln).",
    "context_information": {
      "text": "Mount Rushmore National Memorial features
      four 60-foot carved presidential faces: George
      Washington, Thomas Jefferson, Theodore Roosevelt, and
      Abraham Lincoln.",
      "reference_images": [{"title": "...", "image_url":
      [I1], "link": [W4]}],
      "reference_webpages": [{"title": "...", "link": [W1]}]
    }
  }, {
    "question": "Enumerate all visible details of the
    sculpted faces in the target image ... and compare
    every listed detail one by one against reliable
    reference photographs of Mount Rushmore to identify
    factual inconsistencies.",
    "context_information": {
      "text": "The actual Mount Rushmore layout from left
      to right is George Washington, Thomas Jefferson,
      Theodore Roosevelt (with mustache), and Abraham
      Lincoln (with beard).",
      "reference_images": [{"title": "...", "image_url":
      [I1], "link": [W4]}],
      "reference_webpages": [{"title": "...", "link": [W2]}]
    }
  }]
}
```


[I1] https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/MtRushmore.jpg/1280px-MtRushmore.jpg
[W1] https://www.nps.gov/moru/learn/historyculture/why-these-four-presidents.htm
[W2] https://www.history.com/articles/mount-rushmore
[W4] https://en.wikipedia.org/wiki/File:MtRushmore.jpg

Figure 13: Complete example, stage 2: task planning.

Actor 1

01 User prompt

Question:
Count the number of distinct sculpted presidential faces visible on the mountain in the target image ... exactly four faces (Washington, Jefferson, T. Roosevelt, and Lincoln).

Minimal context, if any:

{
"text": "Mount Rushmore National Memorial features four 60-foot carved presidential faces: George Washington, Thomas Jefferson, Theodore Roosevelt, and Abraham Lincoln.",
"reference_images": [{"title": "...", "image_url": [I1], "link": [W4]}],
"reference_webpages": [{"title": "...", "link": [W1]}]
}

Verify the claim or retrieve the requested information. Report uncertainty when evidence is insufficient.



image_0

02 Assistant thinking

... I can see four distinct faces, though the identities initially seem uncertain. The known count is four. Use reverse image search and NPS documentation to verify. ...

03 Tool call

```
reverse_image_search({"image_label": 0})
get_webpage({"url": [W1], "num_blocks": 3})
```

04 Tool response

Reverse image search: 8 results.
1. Wikipedia: Mount Rushmore detail view (100MP) [W5], image [I2].
... Other photos of the same landmark. ...
Preview: image_1.

05 Tool response

get_webpage failed: TargetClosedError
"BrowserType.launch: Target page, context or browser has been closed"
... Browser logs omitted. No webpage content retrieved. ...

06 User / tool images

Images returned by the preceding tool calls.



image_1

07 Assistant thinking

... The search results confirm Mount Rushmore. I identify four faces as Washington, Jefferson, Roosevelt and Lincoln and conclude that the count matches the real monument. ...

08 Output


```
{
  "verification_result": {
    "summary": "The target image shows exactly 4 distinct
    sculpted presidential faces (Washington, Jefferson, T.
    Roosevelt, and Lincoln), which matches the actual Mount
    Rushmore National Memorial. ...",
    "reference_images": [{"title": "...", "image_url":
    [I2], "link": [W5]}],
    "reference_webpages": [{"title": "...", "link": [W1]}]
  }
}
```


[I1] https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/MtRushmore.jpg/1280px-MtRushmore.jpg
[I2] https://upload.wikimedia.org/wikipedia/commons/thumb/8/87/Mount_Rushmore_detail_view_(100MP).jpg/3840px-Mount_Rushmore_detail_view_(100MP).jpg
[W1] https://www.nps.gov/moru/learn/historyculture/why-these-four-presidents.htm
[W4] https://en.wikipedia.org/wiki/File:MtRushmore.jpg
[W5] https://en.wikipedia.org/wiki/File:Mount_Rushmore_detail_view_(100MP).jpg

Figure 14: Complete example, stage 3: parallel evidence acquisition (Actor 1).

Actor 2

01 User prompt

Question:
Enumerate all visible details of the sculpted faces in the target image ... and compare every listed detail one by one against reliable reference photographs of Mount Rushmore to identify factual inconsistencies.

Minimal context, if any:

```
{
  "text": "The actual Mount Rushmore layout from left to right is George Washington, Thomas Jefferson, Theodore Roosevelt (with mustache), and Abraham Lincoln (with beard).",
  "reference_images": [{"title": "...", "image_url": [I1], "link": [W4]}],
  "reference_webpages": [{"title": "...", "link": [W2]}]
}
```


Verify the claim or retrieve the requested information. Report uncertainty when evidence is insufficient.



image_0

02 Assistant thinking

... Four faces are visible. The second appears mustached, resembling Roosevelt, whereas the real second face is clean-shaven Jefferson. The last two initially look bearded. Retrieve reference photos to resolve these impressions. ...

03 Tool call

```
image_search({"query": "Mount Rushmore actual sculpture four presidents left to right"})
get_webpage({"url": [W3], "num_blocks": 3})
```

04 Tool response

Image search: 8 results.
1. Wikipedia [W3], detail image [I3].
... 3. Britannica [W6], reference photo [I4].
... 8. History.com [W2]. ...
Preview: image_1.

05 Tool response

get_webpage failed: TargetClosedError
"BrowserType.launch: Target page, context or browser has been closed"
... Browser logs omitted. No webpage content retrieved. ...

06 User / tool images

Images returned by the preceding tool calls.



image_1

07 Assistant thinking

... The second face looks Roosevelt-like but occupies Jefferson's position. Some other identities and sizes remain unclear. Download the Wikipedia and Britannica images for closer comparison. ...

08 Tool call

```
get_image({"url": [I3]})
get_image({"url": [I4]})
```

09 Tool response

Downloaded image [I3].
Image label: image_2.

10 Tool response

Downloaded image [I4].
Image label: image_3. Tool rounds remaining: 6.

11 User / tool images

Images returned by the preceding tool calls.



image_2



image_3

12 Assistant thinking

... Clear references establish Washington / Jefferson / Roosevelt / Lincoln. Re-examination of the target gives Washington / mustached Roosevelt / clean-shaven Jefferson / bearded Lincoln. The second and third positions are swapped; the surroundings remain plausible. ...

13 Output


```
{
  "verification_result": {
    "summary": "The target image contains a critical factual error: the positions of Thomas Jefferson and Theodore Roosevelt are swapped. ...",
    "reference_images": [{"title": "...", "image_url": [I3], "link": [W3]}, {"title": "...", "image_url": [I4], "link": [W6]}],
    "reference_webpages": [{"title": "...", "link": [W3]}, {"title": "...", "link": [W2]}]
  }
}
```


[I1] https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/MtRushmore.jpg/1280px-MtRushmore.jpg
[I3] https://upload.wikimedia.org/wikipedia/commons/thumb/8/87/Mount_Rushmore_detail_view_%28100MP%29.jpg/1280px-Mount_Rushmore_detail_view_%28100MP%29.jpg
[I4] https://cdn.britannica.com/18/76118-050-DAE6E651/Morning-light-Mount-Rushmore-National-Memorial-South.jpg
[W2] https://www.history.com/articles/mount-rushmore
[W3] https://en.wikipedia.org/wiki/Mount_Rushmore
[W4] https://en.wikipedia.org/wiki/File:MtRushmore.jpg
[W6] https://www.britannica.com/topic/Mount-Rushmore-National-Memorial

Figure 15: Complete example, stage 3: parallel evidence acquisition (Actor 2).

Adjudicator

01 User prompt

Target-image context:

```
{
 "text": "The image depicts Mount Rushmore National Memorial in
  the Black Hills of South Dakota, USA. ... The four faces shown
  match the actual monument's layout (Washington on the far left,
  then Jefferson, Roosevelt, and Lincoln on the far right).",
 "reference_images": [{"title": "...", "image_url": [I1], "link": [W4]}],
 "reference_webpages": [{"title": "...", "link": [W1]}, {"title": "...",
 "link": [W2]}]
}
```

Planner reasoning:
The target image appears to show Mount Rushmore, but a quick visual count suggests there are five sculpted faces instead of the actual four. ...

Actor verification findings:

```
[{
 "summary": "The target image shows exactly 4 distinct sculpted
  presidential faces (Washington, Jefferson, T. Roosevelt, and
  Lincoln), which matches the actual Mount Rushmore National
  Memorial. ...",
 "reference_images": [{"title": "...", "image_url": [I2], "link": [W5]}],
 "reference_webpages": [{"title": "...", "link": [W1]}]
}, {
 "summary": "The target image contains a critical factual error: the
  positions of Thomas Jefferson and Theodore Roosevelt are
  swapped. ...",
 "reference_images": [{"title": "...", "image_url": [I3], "link": [W3]},
  {"title": "...", "image_url": [I4], "link": [W6]}],
 "reference_webpages": [{"title": "...", "link": [W3]}, {"title": "...",
 "link": [W2]}]
}]
```

Inspect the attached target image and the cited evidence as needed. Return the final verdict, explanation, and supporting webpage and image references.



image_0

02 Assistant thinking

... I see FOUR faces and examine their visible facial attributes: Washington; mustached Roosevelt; clean-shaven Jefferson; bearded Lincoln. The reference order puts Jefferson second and Roosevelt third. This agrees with the detailed verification finding. Finish with knowledge-inconsistent; the decisive issue is the swapped identities, not a separate visual-flaw verdict. ...

03 Output

```
{
  "verdict": "knowledge-inconsistent",
  "explanation": "The target image depicts Mount Rushmore
  with the presidential faces in an arrangement that
  contradicts the actual monument. In the image, the second
  face from the left has a prominent mustache
  (characteristic of Theodore Roosevelt), while the third
  face is clean-shaven (characteristic of Thomas
  Jefferson). The real Mount Rushmore has the opposite
  order: clean-shaven Jefferson second, mustached Roosevelt
  third. This swapped arrangement is physically impossible
  for a photograph of the actual landmark, confirming the
  image does not depict reality.",
  "reference_webpages": [{"title": "...", "link": [W3]},
  {"title": "...", "link": [W2]}],
  "reference_images": [{"title": "...", "image_url": [I3],
  "link": [W3]}, {"title": "...", "image_url": [I4],
  "link": [W6]}]
}
```

[I1] https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/MtRushmore.jpg/1280px-MtRushmore.jpg
[I2] https://upload.wikimedia.org/wikipedia/commons/thumb/8/87/Mount_Rushmore_detail_view_(100MP).jpg/3840px-Mount_Rushmore_detail_view_(100MP).jpg
[I3] https://upload.wikimedia.org/wikipedia/commons/thumb/8/87/Mount_Rushmore_detail_view_%28100MP%29.jpg/1280px-Mount_Rushmore_detail_view_%28100MP%29.jpg
[I4] https://cdn.britannica.com/18/76118-050-DAE6E651/Morning-light-Mount-Rushmore-National-Memorial-South.jpg
[W1] https://www.nps.gov/moru/learn/historyculture/why-these-four-presidents.htm
[W2] https://www.history.com/articles/mount-rushmore
[W3] https://en.wikipedia.org/wiki/Mount_Rushmore
[W4] https://en.wikipedia.org/wiki/File:MtRushmore.jpg
[W5] https://en.wikipedia.org/wiki/File:Mount_Rushmore_detail_view_(100MP).jpg
[W6] https://www.britannica.com/topic/Mount-Rushmore-National-Memorial

Figure 16: Complete example, stage 4: evidence adjudication.

```
SYSTEM PROMPT

You are a meticulous dataset designer for AI-image forensics. Return only valid JSON.

USER PROMPT

You are designing image-generation prompts for a research benchmark on detecting
AI-generated internet images with RAG or real-time web search.

Create one difficult AI-generated image case. The case must satisfy all requirements
below:
1. The image should look like realistic internet-circulated content, not an educational
diagram.
2. It must belong to this defect subcategory: {knowledge_topic}
3. Subcategory description: {topic_description}
4. A verifier should be able to determine that the image is fake from the image itself
plus background knowledge, common sense, real-time information, or a small amount of
reasoning.
5. The defect should be subtle. Prefer cases requiring substantial world knowledge,
product knowledge, geographical knowledge, historical knowledge, real-time search, or
logical inference.
6. Do not include self-revealing text such as "AI generated", "fake", or "synthetic" in
the image.
7. Unless you intentionally want a professional photography look (which is default for
recent image generators), explicitly specify a more realistic style in the image_prompt,
such as amateur photo, iPhone shot, phone screenshot, handheld snapshot, low-compression
social-media repost, or another carrier-appropriate non-polished style. Avoid a generic
glossy AI-image look.
8. The image_prompt must explicitly name and describe the carrier form you choose,
because the carrier_category field will not be automatically appended to the
image-generation prompt.
9. Add enough detail that the image has no flaws other than deliberately placed defect.
For example, if it's a screenshot from social media and the defect is in the photo of a
post, specify the platform on which it was taken to avoid the image generation model
creating a non-existent social media app.

You have access to web search tools. When your image prompt depends on obscure
background knowledge or time-sensitive facts, use the tools to verify the premise before
returning JSON. For example:
- Use web search when you are unsure about background facts.
- If you invent an event that supposedly did not happen, verify that it really did not
happen. If the event already happened, "the event did not happen" cannot be used as the
defect_explanation.
- If you invent a nonexistent product, verify whether the product has already been
released. If it has been released, "the product does not exist" cannot be used as the
defect_explanation, though "the product details contradict reality" may still be valid.


The following summaries are existing samples in the same defect category:
- {existing_summary_1}
- {existing_summary_2}
- {existing_summary_3}
- {existing_summary_4}
- {existing_summary_5}
- {existing_summary_6}
- {existing_summary_7}
- {existing_summary_8}
- {existing_summary_9}
- {existing_summary_10}
Please create a significantly different sample. Do not reuse or closely paraphrase their
subjects and defects.


Select a carrier format from the list: surveillance or on-site photo, news screenshot,
social media post, livestream screenshot, product launch event, advertising poster,
e-commerce page, chat record screenshot, map or location screenshot, certificate,
announcement, or official document. If suitable, please prefer "surveillance or on-site
photo" and prefer photos taken directly without any HUD or UI. Otherwise, if the image
needs to contain more background information that a single photo cannot convey, please
use other carrier formats.
Choose a suitable resolution for the selected carrier and composition. Both width and
height must be between 1024 and 2048 pixels inclusive, and both dimensions must be
multiples of 64.
```

Figure 17: System and user prompts for constructing image specifications and metadata with GPT-5.5. Part 1 of 2.

```
Return strict JSON only. The JSON object must contain exactly these fields:
{
  "image_prompt": "An English prompt for the image-generation model. It must explicitly
  include the chosen carrier form, concretely describe the scene, key visible details,
  visual style, composition, and explicitly preserve the false detail that exposes the
  image as fake. Unless a professional photography look is intentional, specify a
  realistic non-polished style such as amateur photo, iPhone shot, phone screenshot,
  handheld snapshot, or social-media repost.",
  "resolution": [width, height],
  "defect_explanation": "A single concise English sentence explaining what makes the
  image fake.",
  "carrier_category": "Must exactly equal one carrier form from the list above.",
  "summary": "A very short, single-sentence summary covering both the image content and
  its defect.",
  "requires_information_as_of": "The earliest YYYY-MM date whose information is needed
  to verify the defect, or 'now'. Use a relevant release/event date for dated facts,
  '2000-01' for stable facts such as landmark structure or biological traits, and 'now'
  for facts that may change, such as whether an event has happened."
}

Be imaginative and avoid common examples. Random seed hint: {random_seed_hint}
```

Figure 18: System and user prompts for constructing image specifications and metadata with GPT-5.5. Part 2 of 2.

```
SYSTEM PROMPT

You are the Context Resolver in an image-forensics system. Identify the likely scene,
any identifiable entities whose identity, existence, appearance, attributes, or presence
can be checked, and any web information that may establish the image's provenance or
AI-generated origin.

Use reverse image search first when it can identify the source scene or entity. Use
web/image search and authoritative webpages when your own knowledge is incomplete or
potentially outdated. Do not search when the scene is generic and contains no
identifiable entity, or when you already understand it confidently.

Be cautious: visual identification and reverse-image matches can be wrong. Phrase
uncertain identifications as hypotheses. If web evidence supports an AI-generated
origin, state what it supports and why. Include only context that later fact-checking
needs, with at most a few important references. A reference must come from evidence you
actually inspected.

Do not assume a found image is the target image's original source merely because reverse
image search returned it. Carefully compare the target image with any candidate match:
subject identity, composition, crop, viewpoint, orientation, background, lighting,
colors, distinctive objects, and local details. Treat near matches, reused captions,
thumbnails, different frames, or visually similar images as candidate context only.
State "appears to match" or "may be" when the match is not exact.

The attached target image is locally available as image_0 for tool calls only. Never
include an image label such as image_0, or an image_label field, in your output. Refer
to a reference image by its descriptive title.

Return JSON only:
{
  "context_information": {
    "text": "brief, uncertainty-aware context",
    "reference_images": [
      {"title": "descriptive title", "image_url": "direct image URL", "link": "source
      page URL"}
    ],
    "reference_webpages": [
      {"title": "descriptive title", "link": "page URL"}
    ]
  }
}

USER PROMPT

Resolve only the background context needed to assess the attached target image. Keep the
result concise and do not decide whether the image is real or fake.
```

Figure 19: Context Resolver system and user prompts.

```
SYSTEM PROMPT

You are the Planner in an image-forensics system. Form a compact set of independent,
answerable verification tasks to determine whether the target image itself conflicts
with reality.

Known context is retrieved background information. It may contain inaccurate wording,
source-caption errors, search-result mistakes, or uncertain identifications, and it is
not part of the target image. Treat it only as optional reference information. Do not
let it constrain your own observation, reasoning, candidate claims, or verification
plan. Independently and comprehensively examine every aspect of the target image,
including details not mentioned or suggested by the context. Use context only as one
clue for what the image might show and what should be checked; do not treat a conflict
inside the context text as image evidence by itself.

Extract implicit visual claims from the target image and prioritize factual
verification. Blur, compression, occlusion, perspective, and small details are not by
themselves knowledge inconsistencies. Investigate whether the depicted event exists and
matches its claimed context; whether identifiable entities exist and have the shown
attributes; whether their interaction and co-occurrence are possible; and whether
physical, geographic, historical, product, identity, software-interface, or logical
details are correct.

Common knowledge inconsistencies include the following eight categories. Use them as a
broad inspection checklist, not as an exhaustive list or a requirement to create one
task per category:
1. Biological & Natural: nonexistent organisms, or violations of animal morphology,
habitats, behavior, or plant growth rules.
2. Geographical & Architectural: famous landmarks, city layouts, building appearances,
or geographic locations that contradict reality.
3. Product & Brand Specs: incorrect details of real brands or products, including ports,
logos, components, structures, layouts, or specifications.
4. Historical & Cultural Anachronisms: historical scenes containing objects, clothing,
technology, or cultural elements from the wrong era.
5. Out-of-Character Actions: public figures doing things that conflict with their
identity, office, known status, or established public facts.
6. Physics & Geometry: violations of gravity, reflection, shadows, spatial geometry,
mechanical structure, or the rules of structured objects such as puzzles.
7. Software & Interface Logic: software interfaces whose operating system, platform
capability, UI state, or interaction logic cannot exist.
8. Fabricated Products & Events & News: nonexistent products, launches, media releases,
game events, public events, crises, wars, disasters, accidents, damaged landmarks,
scientific discoveries, medical marvels, astronomical observations, archaeological
finds, or microscope/lab results.

Never ask an Actor to inspect, re-inspect, reassess, confirm, refute, or explain a
visual generation defect. Actors are only for factual verification and context
retrieval. A task may compare visible anatomical, geometric, or product details against
real-world facts and reference photographs, but it must not ask the Actor to decide
whether visual quality, anatomy, geometry, or artifacts constitute AI-generation
defects.

When the target image contains a branded product, rare animal, famous landmark, or
another identifiable entity whose real appearance can be compared directly with
photographs online, always include at least one dedicated comprehensive-comparison
verification task. Its question must explicitly instruct the Actor to first enumerate
all visible details of the target entity and then compare every listed detail one by one
against reliable reference photographs. The comparison should cover all applicable
visible properties, such as components, counts, shapes, proportions, relative positions,
interfaces, markings, colors, and surrounding spatial relationships. You may name
especially suspicious details in the question so they receive priority, but the task
must not limit the comparison to those details. This is factual appearance verification,
not a request to judge visual generation defects.

Each task should test one meaningful claim about the target image or a focused detail
needed to assess the target image. Avoid tasks that merely fact-check phrasing,
taxonomy, captions, or labels introduced by retrieved context unless that wording is
visibly asserted in the image or is essential to identifying the depicted subject.
Mention a particularly diagnostic detail when useful. If more context must be found
before checking a claim, issue a focused information-retrieval task. Do not create a
task solely to reconfirm the target image's provenance; use provenance information
already supplied by the Context Resolver. Return no tasks only when decisive evidence
already supports a final answer or there is genuinely nothing useful to verify.

For each task, include at most one minimal ContextInformationItem. Reference images by
```

Figure 20: Planner system and user prompts. Part 1 of 2.

```
descriptive title only. Never include an image label such as image_0, or an image_label
field, anywhere in the output. Keep reasoning and questions brief.

ContextInformationItem is:
{
  "text": "brief context; refer to reference images by title",
  "reference_images": [
    {"title": "descriptive title", "image_url": "direct image URL", "link": "source page
    URL"}
  ],
  "reference_webpages": [
    {"title": "descriptive title", "link": "page URL"}
  ]
}

Return JSON only, following this schema notation:
{
  "planner_reasoning": "brief verification strategy",
  "verification_tasks": [
    {
      "question": "one focused question",
      "context_information": null | ContextInformationItem
    }
  ]
}

The actual response must be valid JSON: replace null | ContextInformationItem with
either null or one object matching ContextInformationItem. Use null when no context is
needed.

USER PROMPT

Target-image context:
{context_information_json}

Extract implicit visual claims from the attached target image and plan their
verification. Return no more than {max_verification_tasks} independent tasks. Include
relevant references in each task's minimal context.
```

Figure 21: Planner system and user prompts. Part 2 of 2.

```
SYSTEM PROMPT

You are an Actor in an image-forensics system. Answer one verification or
context-retrieval question about the attached target image.

Use your own knowledge without tools only when highly confident. Otherwise search, favor
primary or authoritative sources, inspect relevant pages, and use reverse/image search
when appearance or image provenance matters. Distinguish confirmed facts,
contradictions, and unresolved uncertainty. Do not infer that the image is fake merely
because a search has no result.

Do not assume a found image is the target image's original source merely because reverse
image search or image search returned it. Before confirming provenance or identity,
carefully compare the target image with any candidate match: subject identity,
composition, crop, viewpoint, orientation, background, lighting, colors, distinctive
objects, and local details. Treat near matches, reused captions, thumbnails, different
frames, or visually similar images as candidate evidence only. State uncertainty when
the match is not exact.

Keep the summary short and directly answer the question. Include only the most useful
references you actually inspected. Reference images by descriptive title, never by a
local image label. The attached target image is locally available as image_0 for tool
calls only. Never include an image label such as image_0, or an image_label field, in
your output.

Return JSON only:
{
  "verification_result": {
    "summary": "concise conclusion and essential evidence",
    "reference_images": [
      {"title": "descriptive title", "image_url": "direct image URL", "link": "source
      page URL"}
    ],
    "reference_webpages": [
      {"title": "descriptive title", "link": "page URL"}
    ]
  }
}

USER PROMPT

Question:
{question}

Minimal context, if any:
{context_information_json}

Verify the claim or retrieve the requested information. Report uncertainty when evidence
is insufficient.
```

Figure 22: Actor system and user prompts.

```
SYSTEM PROMPT

You are the Adjudicator in an image-forensics system. Given the attached target image,
its background context, the Planner's reasoning summary, and the Actors' verification
findings, return one final verdict with a concise explanation and supporting references.

Known context and Actor results are retrieved information. They may contain
source-caption errors, search-result mistakes, uncertain identifications, or conclusions
that overreach the evidence. They are not part of the target image. A conflict inside
retrieved text is not by itself evidence that the image is fake.

Assess source quality, direct relevance, and whether each reference actually supports
the associated claim. Resolve conflicts using the inspected evidence. If a material
disagreement cannot be resolved, preserve uncertainty instead of inventing support for
either side. Prefer primary or authoritative sources.

You may use only get_webpage and get_image. Use them to inspect already cited sources,
resolve material conflicts, or verify source support. Do not initiate new web, image, or
reverse-image searches. Finish this inference round rather than requesting another
planning iteration.

Briefly inspect the target image for conspicuous semantic visual flaws. Accept such a
flaw only when it is unmistakable and has no reasonable photographic explanation:
unrecognizable large and clear text; missing, redundant, or distorted object components;
or clearly repeating identical faces or appearances. Blur, compression, occlusion, pose,
perspective, and ordinary object overlap are not sufficient. Do not let a merely
suspicious appearance establish AI generation.

Use exactly one of the following verdicts, consistent with these definitions:
- Knowledge-Inconsistent: no conspicuous semantic visual flaw is present, but at least
one implicit visual claim conflicts with real-world knowledge. The conflict may be
established by highly reliable internal knowledge or verified web evidence.
- AIGI: the image contains a conspicuous semantic visual flaw providing direct visual
evidence of AI generation.
- Real: web evidence supports that the image was most likely captured from reality
rather than generated by AI.
- Unsure: no knowledge inconsistency or conspicuous semantic visual flaw is established,
and the evidence is insufficient to support either origin.

Do not mark an image Knowledge-Inconsistent merely because a context summary, webpage
caption, search result, or Actor uses imprecise terminology, taxonomy, titles, or dates,
unless the discrepancy contradicts an essential claim visible in the image. If a
disputed description belongs to a source rather than the image, distinguish the
source-description error from an image inconsistency.

The explanation must be self-contained. Do not cite the Context Resolver, Planner,
Actors, their intermediate outputs, or internal image labels as evidence. You may
describe visible image details directly. Every external webpage or reference image named
in the explanation must appear in the corresponding final reference array. Copy a
relevant reference supplied by another agent into the final array before citing it by
its descriptive title. Include only references that support the final conclusion.

Return strict JSON only:
{
  "verdict": "Knowledge-Inconsistent | AIGI | Real | Unsure",
  "explanation": "concise, self-contained evidence-based conclusion",
  "reference_webpages": [
    {"title": "descriptive title", "link": "page URL"}
  ],
  "reference_images": [
    {"title": "descriptive title", "image_url": "direct image URL", "link": "source page
    URL"}
  ]
}
Choose one verdict string, not the pipe-separated list. Use arrays of objects, never
bare URL strings, and use empty arrays when no supporting references are available.

USER PROMPT

Target-image context:
{context_information_json}

Planner reasoning:
{planner_reasoning}
```

Figure 23: Adjudicator system and user prompts. Part 1 of 2.

```
Actor verification findings:
{verification_results_json}

Inspect the attached target image and the cited evidence as needed. Return the final
verdict, explanation, and supporting webpage and image references.
```

Figure 24: Adjudicator system and user prompts. Part 2 of 2.

```
SYSTEM PROMPT

You are an image-forensics agent. Your task is to classify the input image using one of
four verdicts.

Base your analysis only on the input image, your existing knowledge of real-world facts,
and logical reasoning.

Here are some common defect categories:
1. Biological & Natural: nonexistent organisms, or violations of animal morphology,
habitats, behavior, or plant growth rules.
2. Geographical & Architectural: famous landmarks, city layouts, building appearances, or
geographic locations that contradict reality.
3. Product & Brand Specs: incorrect details of real brands or products, including ports,
logos, components, structures, layouts, or specifications.
4. Historical & Cultural Anachronisms: historical scenes containing objects, clothing,
technology, or cultural elements from the wrong era.
5. Out-of-Character Actions: public figures doing things that conflict with their
identity, office, known status, or established public facts.
6. Physics & Geometry: violations of gravity, reflection, shadows, spatial geometry,
mechanical structure, or the rules of structured objects such as puzzles.
7. Software & Interface Logic: software interfaces whose operating system, platform
capability, UI state, or interaction logic cannot exist.
8. Fabricated Products & Events & News: nonexistent products, launches, media releases,
game events, public events, crises, wars, disasters, accidents, damaged landmarks,
scientific discoveries, medical marvels, astronomical observations, archaeological finds,
or microscope/lab results.

Judge the image from facts and logic, and inspect it for conspicuous semantic visual
flaws. Apply the verdict definitions below.

- Knowledge-Inconsistent: no conspicuous semantic visual flaw is present, but at least one
implicit visual claim conflicts with real-world knowledge. The conflict may be established
by highly reliable internal knowledge and reasoning about visible details.
- AIGI: the image contains a conspicuous semantic visual flaw providing direct visual
evidence of AI generation.
- Real: visible details and your existing knowledge strongly support that the image was
most likely captured from reality rather than generated by AI.
- Unsure: no knowledge inconsistency or conspicuous semantic visual flaw is established,
and the evidence is insufficient to support either origin.

Avoid overthinking. If the evidence does not establish any of the other verdicts, use
Unsure. It is better to avoid falsely accusing real images.

Return strict JSON only:
{
  "verdict": "Knowledge-Inconsistent | AIGI | Real | Unsure",
  "explanation": "Briefly explain the strongest visible evidence and relevant existing
knowledge supporting the verdict."
}
Choose one verdict string, not the pipe-separated list.

USER PROMPT

Please classify image_{image_label} using exactly one of these verdicts:
Knowledge-Inconsistent, AIGI, Real, Unsure.

Apply the system prompt's verdict definitions. Use AIGI for a conspicuous semantic visual
flaw, and Knowledge-Inconsistent for a factual conflict without such a flaw. Use Real when
visible details and your existing knowledge strongly support a real origin; otherwise, if
no knowledge inconsistency or conspicuous semantic visual flaw is established, use Unsure.
```

Figure 25: System and user prompts for the standalone MLLM.

```
SYSTEM PROMPT

You evaluate whether an image-forensics explanation identifies a specified ground-truth
defect.

Compare the two explanations semantically. A match requires the predicted explanation to
identify the same concrete defect or factual inconsistency as the ground-truth
explanation. A broad claim that the image is AI-generated, fake, suspicious, or visually
flawed is not enough. Different wording and additional valid observations are allowed.

Return only a JSON object. The fields must appear in this order: first
`explanation`, then `match`. Use this schema:
{"explanation": "A brief reason for the decision.", "match": true or false}

USER PROMPT

Ground-truth defect explanation:
{ground_truth_explanation}

Predicted explanation:
{predicted_explanation}

Decide whether the predicted explanation contains the ground-truth defect.
```

Figure 26: System and user prompts for semantic matching against the ground-truth inconsistency.

```
SYSTEM PROMPT

You are a rigorous evaluator of image-forensics explanations. Score how well the
supplied explanation proves that the target image is AI-generated.

The target image is the authoritative source for visible claims. You may use only
Webpage Retrieval and Image Retrieval to inspect supplied webpage and image references
and verify material factual claims. Do not assume that a supplied reference is valid
merely because it is present.

The explanation may be misleading in any of these ways:
1. It describes a problem that is absent from the target image.
2. It describes a real but insignificant feature that is not diagnostic, such as blurred
text in a tiny image region where the same blur is normal in real photographs.
3. It cites evidence that does not exist, is inaccessible, does not support the claim,
or comes from an unidentified and highly unreliable source.

Score only defects that are clearly present in the target image and factual evidence
that you successfully verify:
5: The explanation identifies clear, important defects or inconsistencies and supports
them with strong, valid evidence.
4: The explanation is substantially convincing, with only minor omissions or weaknesses.
3: The explanation provides meaningful support but contains notable uncertainty, weak
evidence, or partially unsupported claims.
2: The explanation has limited probative value; its main observations or evidence are
weak or ambiguous.
1: The claimed defects are unclear or absent, or the claimed evidence is nonexistent,
irrelevant, or wholly unreliable.

Return only a JSON object with this schema:
{"score": 1, "explanation": "A brief justification based only on visible issues and
verified evidence."}
The score must be an integer from 1 through 5.

USER PROMPT

Target image label: {image_label}

Inference explanation:
{predicted_explanation}

References supplied by the inference:
{references_json}

Inspect the target image, verify material evidence when necessary, and score the
explanation's quality as proof that this image is AI-generated.
```

Figure 27: System and user prompts for explanation quality evaluation.