ClueAegis: Heuristic-to-Reasoning Cognitive-skill Learning for Unified Evidence-based Synthetic Image Detection

Huangsen Cao[1,*], Hongkang Chu[3], Yuxi Li[2], Ying Zhang[2], Chen Li[2], Jing LYU[2], Yongwei Wang[1,✉], Yu Zhao[1], Fei Wu[1]
[1]Zhejiang University
[2]WeChat Vision, Tencent Inc
[3]University of the Chinese Academy of Sciences
✉ Corresponding author: yongwei.wang@zju.edu.cn

Abstract

The rapid advancement of generative models has made synthetic images increasingly realistic, challenging reliable detection. Existing methods are often limited to end-to-end classification or monolithic reasoning, and thus fail to model structured forensic reasoning and heterogeneous visual evidence. We revisit synthetic image detection from a cognitive perspective and propose a *Heuristic-to-Reasoning* cognitive skill learning framework for evidence-based forensic analysis. Given an input image, our framework first extracts heuristic perceptual clues, selects the optimal forensic skill, and then performs skill-conditioned reasoning for evidence extraction and decision making. To support this paradigm, we introduce **ClueAegis-Bench**, which decomposes synthetic image detection into explicitly annotated forensic cognitive skills for structured evaluation beyond binary classification. Based on this benchmark, we propose **ClueAegis** (Cognitive-skill Learning for Unified Evidence-based Synthetic Image Detection), a two-stage agentic framework that conducts heuristic skill selection followed by evidence-guided reasoning through skill-conditioned toolchains. This design reformulates synthetic image detection as a configurable multi-skill reasoning process that bridges perception, skill selection, and forensic reasoning. Extensive experiments show that ClueAegis achieves state-of-the-art performance while improving cross-domain generalization and robustness. It also provides transparent reasoning trajectories and structured forensic evidence, offering a more explainable alternative to conventional end-to-end detectors.

1 Introduction

With the rapid advancement of generative artificial intelligence(Esser et al., 2024; Goodfellow et al., 2020; Ho et al., 2020), high-quality AI-generated images have become increasingly prevalent across online platforms. While these images enrich visual content and enhance expressive diversity, they also blur the boundary between real and synthetic data, potentially facilitating misinformation dissemination and raising security concerns. Therefore, it is crucial to develop effective synthetic image detection methods that can fully leverage visual evidence for reliable discrimination.

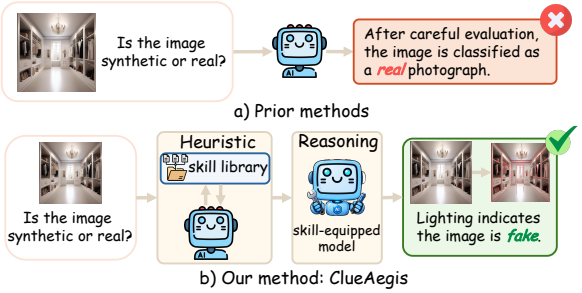


Figure 1: (a) Prior synthetic image detection methods; (b) Our method: ClueAegis with a cognition-guided reasoning paradigm.

Recent advances in synthetic image detection (Chai et al., 2020; Liu et al., 2024; Ojha et al., 2023; Wang et al., 2023) have achieved strong progress; however, most methods still rely on unified end-to-end or monolithic reasoning pipelines. While recent LLM-based approaches (Li et al., 2025a; Cao et al., 2025; Zhou et al., 2025b; Xu et al., 2025; Huang et al., 2025; Li et al., 2025b; Ye et al., 2025) introduce chain-of-thought reasoning for improved interpretability, they often treat heterogeneous forensic evidence as a single undifferentiated signal. However, forensic cues such as lighting, frequency artifacts, OCR errors, and anatomical inconsistencies exhibit fundamentally different distributions, and forcing them into a unified reasoning space can lead to feature entanglement and suboptimal evidence extraction. This suggests that the key limitation lies not only in reasoning depth, but also in the lack of structured

*Work done during an internship at WeChat Vision, Tencent Inc.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14941–14972
October 24-29, 2026 ©2026 Association for Computational Linguistics

decomposition of forensic cognition. Motivated by this, we propose a skill-driven decomposition with a heuristic-to-reasoning mechanism, which separates heterogeneous cues into structured skill spaces before performing higher-level reasoning (see Fig. 1). This enables more faithful modeling of forensic workflows and improves both robustness and generalization.

One fundamental reason why existing methods struggle to learn such decomposed forensic cognition is the absence of heuristic-structured datasets and learning paradigms. Existing datasets (Lu et al., 2023; Zhu et al., 2023; Jiang et al., 2025) are typically organized according to generator types or data sources, rather than structured forensic heuristics and reasoning cues. Although several explainability benchmarks provide artifact-level descriptions, they fail to establish explicit correspondences between forensic evidence and reasoning processes, making it difficult to learn a decomposed heuristic-to-reasoning paradigm.

Therefore, the core limitations of existing methods can be summarized into two key issues: (i) **the lack of sufficiently large-scale and well-annotated datasets for fine-grained skill classification**, which hinders the effective learning of decomposed forensic cues and discriminative forensic skill representations; and (ii) **the absence of an explicit and principled heuristic-to-reasoning modeling paradigm**, which prevents models from adaptively transforming low-level perceptual forensic cues into conditional high-level reasoning processes during inference. Together, these limitations fundamentally restrict existing methods from achieving hierarchical cognitive modeling from fast heuristic perception to slow analytical reasoning.

To address these limitations, we introduce **ClueAegis-Bench**, a cognition-aware benchmark for synthetic image detection with explicit skill decomposition. Unlike conventional binary benchmarks, it adopts a skill-centric annotation scheme grounded in forensic reasoning, where each image is labeled with one of 12 predefined detection skills corresponding to its most discriminative forensic cue. This organizes heterogeneous artifacts into a unified skill space. Based on this decomposition, we reformulate detection as a structured reasoning process that explicitly separates perception, skill selection, and decision making. Models are required to perform perception-based analysis to select the appropriate forensic skill before making the final prediction. ClueAegis-Bench thus provides a testbed for studying skill-aware reasoning and cognitive decomposition in synthetic image detection, thereby improving generalization performance.

Based on ClueAegis-Bench, we propose **ClueAegis**, a two-stage heuristic-to-reasoning cognitive skill selection agent for synthetic image detection. In the first stage, the model operates as a perception-driven agent that interprets the input image and selects the most appropriate detection skill from a predefined skill space, thereby transforming cognitive understanding into an explicit actionable plan for downstream reasoning. In the second stage, the model performs skill-conditioned agentic detection by dynamically invoking corresponding analysis tools and feature transformation operators based on the selected skill. Each skill is associated with a distinct toolchain and processing paradigm, enabling an on-demand execution mechanism for structured forensic reasoning. This formulation reframes conventional static binary classification as a configurable, executable, and multi-skill agentic reasoning framework, thereby improving both adaptability and generalization capability in synthetic image detection.

Our contributions are summarized as follows:

- **Task Re-formulation and Benchmark.** We reformulate synthetic image detection from binary classification to a skill-driven structured reasoning problem, and introduce ClueAegis-Bench to support this new paradigm with skill-oriented annotations.
- **Skill-driven Classification Space.** We organize forensic cues into 12 skill-oriented categories corresponding to different reasoning capabilities, forming a skill-driven classification space for adaptive forensic analysis.
- **Two-stage Forensic Reasoning.** We design a heuristic-to-reasoning two-stage framework that first builds global perceptual understanding and then performs cognition over relevant forensic skills for structured reasoning.
- **Empirical Results.** Experiments on multiple benchmarks show that our method achieves state-of-the-art performance, with improved generalization and robustness.

2 Related Work

2.1 Synthetic Image Detection

Early synthetic image detection methods mainly relied on CNN-based classifiers(Mo et al., 2018;

Nguyen et al., 2019), leveraging artifacts such as checkerboard patterns(Zhang et al., 2019) and frequency-domain traces(Tan et al., 2024a). With the emergence of pretrained vision models, CLIP-based approaches, such as UnivFD(Ojha et al., 2023), achieved strong cross-domain generalization, while subsequent works further improved robustness through architectural and training refinements(Zhang et al., 2025; Liu et al., 2024; Chen et al., 2025b). Other methods, including NPR(Tan et al., 2024b), HyperDet(Cao et al., 2024), AIDE(Yan et al., 2024) and Patchcraft(Zhong et al., 2023) enhanced detection by modeling pixel relationships and synthetic forensic traces. More recently, multimodal large language models have shifted the field toward reasoning-driven detection. Frameworks such as AIGI-Holmes(Zhou et al., 2025b), REVEAL(Cao et al., 2025), Veritas(Tan et al., 2025), and Ivy-Fake(Jiang et al., 2025) introduced interpretable and reasoning-based paradigms for generalized synthetic image forensic analysis.

2.2 MLLM-Based Multi-Agent Frameworks

Recent advances in large language models have enabled agentic frameworks that integrate task decomposition, tool invocation, and multi-step reasoning. Early multimodal agents, such as Visual ChatGPT(Wu et al., 2023), MM-ReAct(Yang et al., 2023), and HuggingGPT(Shen et al., 2023), employed LLMs as planners for coordinating vision models and external tools, while frameworks like AutoGen(Wu et al., 2024) and MetaGPT(Hong et al., 2024) further explored workflow-level agent collaboration. Recent MLLM studies extended this paradigm toward interleaved perception, reasoning, and action through iterative visual reasoning and tool-augmented interaction(Zheng et al., 2025; Hong et al., 2025; Yu et al., 2025; Zhou et al., 2025a). Meanwhile, skill-oriented agent learning methods, such as SkilRL(Xia et al., 2026) and Skill0(Lu et al., 2026), demonstrated that reusable cognitive skills can effectively bridge perception and decision making.

However, existing agentic MLLM frameworks mainly focus on general reasoning or tool use, while synthetic image detection lacks a structured forensic skill taxonomy and reasoning paradigm. ClueAegis addresses this by reformulating detection as a *Heuristic-to-Reasoning* process, where heuristic-guided skill selection identifies forensic strategies, followed by evidence-based agentic reasoning for decision making.

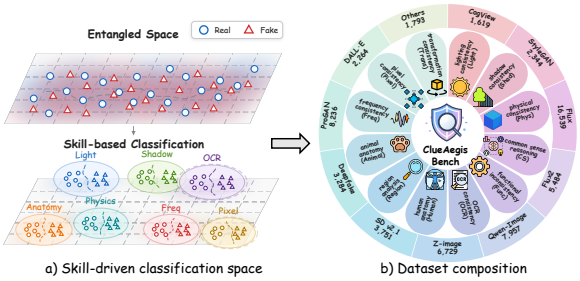


Figure 2: Overview of ClueAegis-Bench. (a) Images are mapped from an entangled real/fake space into skill-specific classification subspaces. (b) Skill-wise partitions across heterogeneous sources form ClueAegis-Bench with diverse synthetic distributions.

3 ClueAegis-Bench

In cognitive psychology, dual-process theory distinguishes System 1—heuristic, intuitive perception—and System 2—deliberate analytical reasoning (Kahneman, 2011). Building on this framework, we formulate synthetic-image detection as a structured forensic task, where System 1 performs rapid heuristic perception and System 2 conducts skill-grounded, evidence-driven reasoning.

3.1 ClueAegis-Bench Overview

To support this paradigm, we construct a structured benchmark for skill-oriented synthetic image detection. As illustrated in Figure 2, the task is formulated as a two-stage classification process: each image is first assigned to one of 12 predefined forensic skill categories, and then used to perform forensic authenticity discrimination conditioned on the predicted skill.

Data Collection and Skill Annotation. We construct a large-scale dataset by aggregating approximately 6M images from multiple public sources, including Fake2M (Lu et al., 2023), Projective-Geometry (Sarkar et al., 2024), TGIF dataset (Mareen et al., 2024), TextAtlas5M (Wang et al.), and WHOOPS (Bitton-Guetta et al., 2023). To further enhance the diversity of generative artifacts, we additionally synthesize 120K images using state-of-the-art generative models, including FLUX (Labs, 2024), FLUX2 (Labs, 2025), Qwen-Image (Wu et al., 2025), and Z-Image (Cai et al., 2025). After data curation and balancing, we obtain a final benchmark consisting of 120K images in total, including balanced real and synthetic samples. Each forensic skill contains approximately 5K real and 5K synthetic images (See *Appendix A*).

For skill annotation, we define a set of 12 human-designed forensic skills that capture complemen-

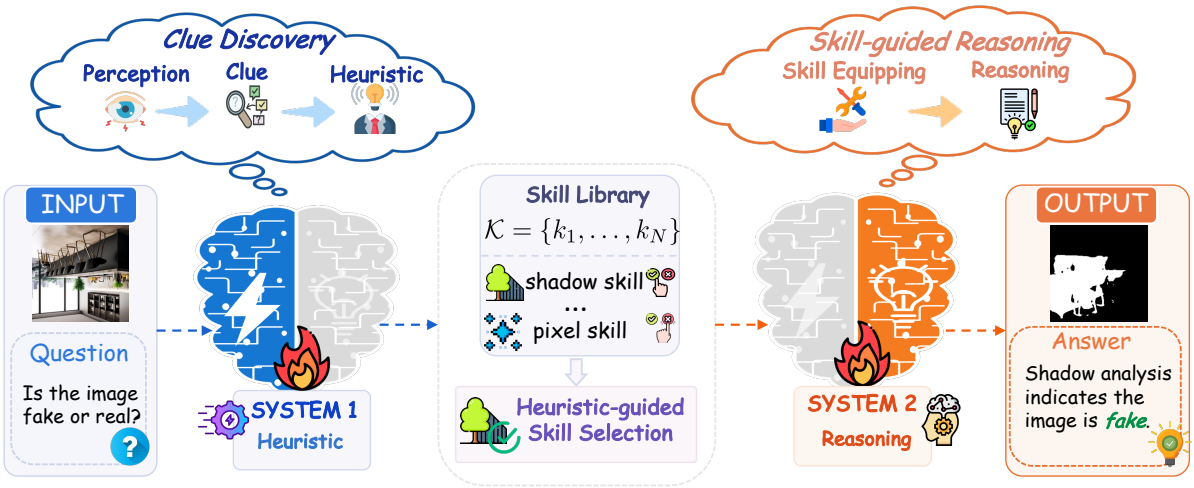


Figure 3: **Overview of ClueAegis.** A two-stage pipeline consisting of System 1 (heuristic) and System 2 (reasoning).

tary types of authenticity cues, including *lighting consistency (Light), shadow consistency (Shad), physical consistency (Phys), common-sense reasoning (CS), functional inconsistency (Func), OCR consistency (OCR), human anatomy (Human), region analysis (Region), animal anatomy (Animal), frequency consistency (Freq), pixel consistency (Pixel), and transformation consistency (Trans)*. Given a synthetic image, we employ **Qwen3-VL-235B-A22B** (Bai et al., 2025) as an automatic multimodal forensic reasoner to evaluate whether each predefined skill can successfully detect forgery evidence in the image. Each skill is treated as an independent forensic detector. If Qwen3-VL determines that a specific skill reveals discriminative cues indicating that the image is synthetic, we retain the corresponding (image, skill) pair as a valid annotation. In this way, the dataset is constructed by linking fake images with the forensic skills that are most effective at exposing their artifacts, rather than performing conventional image-to-label assignment. For skills requiring external grounding signals, we integrate specialized tools; for example, OCR consistency is supported by PaddleOCR (Cui et al., 2025) for textual verification, among other task-specific tools (detailed in *Appendix G*). Further details on annotation protocols and skill definitions are provided in *Appendix B and G*.

4 Method

In this section, we present the heuristic–to-reasoning framework of ClueAegis.

4.1 Overview of ClueAegis

Inspired by the dual-process theory of human cognition, we propose ClueAegis, a synthetic image detection framework based on a heuristic-to-reasoning cognitive paradigm. As illustrated in Figure 3, ClueAegis reformulates conventional end-to-end detection as a cognitive forensic reasoning process with two stages: **Heuristic-guided Skill Selection** and **Skill-guided Forensic Reasoning**.

Heuristic-guided Skill Selection. Traditional synthetic image detection methods typically rely on unified feature representations and a single-stage binary classification framework. However, forgery artifacts exhibit substantial heterogeneity across semantic domains, physical structures, and generation processes, resulting in highly entangled forensic distributions. Such diversity makes globally shared decision boundaries difficult to generalize across real-world synthetic images.

To address this issue, we reformulate synthetic image detection as a skill-conditioned forensic reasoning problem by decomposing authenticity analysis into multiple forensic cognitive skills. Based on this formulation, we construct a skill-conditioned space for adaptive reasoning.

Given an input image x, perceptual forensic evidence is first extracted through a visual encoder.

$$c = f_\theta(x), \tag{1}$$

where f_θ denotes the cognitive perception encoder and c represents forensic clues composed of abnormal visual patterns, structural inconsistencies, and latent semantic conflicts.

We define a forensic cognitive skill library:

$$\mathcal{K} = \{k_1, k_2, \ldots, k_N\}, \tag{2}$$

where each skill corresponds to a specialized forensic reasoning capability, such as OCR verification,

frequency analysis, physical consistency reasoning, or anatomical analysis, among others (see *Appendix G* for full definitions).

To determine the most relevant forensic reasoning pathway, the model selects a skill from the predefined skill library conditioned on the extracted forensic evidence:

$$s^* = \text{VLLM-Select}(c, \mathcal{K}), \tag{3}$$

where selection is performed by prompting a large language model to match forensic evidence with the most relevant cognitive skill in the skill library.

The selected skill then guides subsequent forensic reasoning for authenticity prediction, forming a skill-conditioned inference process, which can be viewed as a System 1-like heuristic routing stage that rapidly identifies the most relevant forensic skill.

Skill-guided Forensic Reasoning. After selecting the forensic skill s^*, ClueAegis performs skill-guided forensic reasoning for evidence-driven authenticity analysis. Specifically, the selected skill is transformed into a specialized forensic prompt that guides the large language model to focus on corresponding anomaly patterns and forensic cues.

Conditioned on the selected skill, the reasoning process is formulated as:

$$\mathcal{R} = P_{\mathcal{L}}(x, s^*), \tag{4}$$

where $P_{\mathcal{L}}$ denotes the skill-conditioned forensic reasoning process and $\mathcal{R}$ represents the generated forensic reasoning trajectory.

During reasoning, the model may invoke external forensic tools and APIs for specialized evidence verification:

$$o = \mathcal{A}(x, s^*, \mathcal{R}), \tag{5}$$

where $\mathcal{A}$ denotes the forensic agent system and o represents auxiliary forensic observations returned by external tools.

The final authenticity prediction is obtained by integrating the reasoning trajectory and external forensic evidence:

$$y = \Phi(\mathcal{R}, o), \tag{6}$$

where $\Phi(\cdot)$ denotes the forensic decision function that maps the reasoning trajectory and external evidence to a binary authenticity prediction.

From a cognitive perspective, this process resembles the analytical reasoning mechanism of System 2, where humans perform deliberate evidence verification and progressively refine forensic judgments under task-specific cognitive skills.

Overall, previous methods typically model synthetic image detection as $p(y \mid x, \mathcal{E})$, where $\mathcal{E} = \{e_i\}_{i=1}^{N}$ denotes a set of entangled forensic evidence and prediction is made through joint reasoning over all cues. In contrast, ClueAegis reformulates the problem as $p(y \mid x, e^*)$, where e^* is the most relevant skill-induced forensic clue selected from a predefined set of cognitive skills, enabling focused reasoning on a single discriminative evidence source.

4.2 Two-Stage Cognitive Training

Our two-stage design models the cognitive inference process rather than the optimization pipeline. Specifically, ClueAegis decomposes synthetic image detection into two sequential stages: (1) heuristic-guided forensic skill selection, and (2) skill-guided reasoning. To learn this behavior, we employ supervised fine-tuning (SFT) followed by group relative policy optimization (GRPO).

Stage I: Heuristic Skill Learning. In the first stage, the model learns to analyze forensic evidence and select the most relevant forensic skill for a given image. Given an input image x, the model predicts the corresponding forensic skill s^*:

$$\mathcal{L}_{\text{skill}} = -\log P_\theta(s^* \mid x). \tag{7}$$

This stage establishes the mapping between heuristic forensic evidence and the predefined forensic skill space.

Stage II: Skill-guided Forensic Reasoning. In the second stage, the model performs authenticity analysis conditioned on the selected forensic skill. Given x and s^*, the model conducts skill-specific reasoning and predicts the final authenticity label y:

$$\mathcal{L}_{\text{det}} = -\log P_\theta(y \mid x, s^*). \tag{8}$$

This stage enables the model to adapt its forensic reasoning behavior according to different cognitive skills.

GRPO-based Cognitive Alignment. To improve reasoning consistency and robustness, we apply GRPO for preference optimization over forensic outputs. The reward function jointly considers answer correctness, skill selection consistency, and format compliance:

$$R = \lambda_1 R_{\text{ans}} + \lambda_2 R_{\text{skill}} + \lambda_3 R_{\text{format}}. \tag{9}$$

Method	Forensic Consistency Metrics												Avg.
	Light	Shadow	Phys	CS	Func	OCR	Human	Region	Animal	Freq	Pixel	Trans	
CNNSpot	94.15	93.00	93.55	94.35	94.30	93.15	93.50	63.80	93.85	64.00	93.00	85.80	88.04
FreqNet	80.10	79.50	89.95	93.70	92.50	87.15	92.70	58.50	90.60	87.10	88.40	74.20	84.53
UnivFD	93.50	87.25	90.90	92.70	92.00	90.35	92.70	65.15	92.10	87.30	82.85	75.20	86.83
NPR	82.60	75.70	87.20	88.75	88.20	82.70	87.40	81.95	87.50	88.40	85.20	82.70	84.86
HyperDet	50.30	50.10	86.90	89.85	86.60	50.30	90.60	50.20	90.35	86.75	80.25	68.80	73.42
AIDE	93.80	90.65	94.30	94.55	94.30	93.00	94.55	64.85	94.25	93.50	92.70	83.45	90.33
DDA	94.00	92.70	94.55	94.65	94.25	93.35	94.60	73.10	94.60	94.35	93.60	90.45	92.02
AIGI-Holmes	51.75	51.05	87.75	91.15	83.40	91.15	91.45	84.85	83.50	98.95	97.75	83.40	83.01
REVEAL	87.25	85.35	80.55	91.25	91.30	92.60	91.55	80.45	93.65	74.45	72.95	61.35	83.56
Veritas	88.35	82.45	76.35	81.70	76.05	63.85	86.75	71.25	80.25	79.60	75.15	65.60	77.28
Ivy-xDetector	50.65	50.95	92.95	94.75	91.25	72.95	94.70	80.55	94.95	78.65	89.75	64.30	79.70
ClueAegis	**99.80**	**99.70**	**99.40**	**99.75**	**99.75**	**99.85**	**99.70**	**95.50**	**99.75**	**99.95**	**99.55**	**95.40**	**99.01**

Table 1: **Performance on ClueAegis-Bench.** ClueAegis achieves the best overall performance and surpasses the previous state-of-the-art method by **6.99%** in accuracy, demonstrating superior generalization capability across diverse synthetic image distributions. **Best** and second-best are highlighted.

Here, R_{ans} measures prediction correctness against ground truth, R_{skill} checks whether the selected forensic skill matches the annotation, and R_{format} enforces the required two-stage reasoning structure (skill selection followed by prediction).

Given sampled outputs $\{o_1, \ldots, o_G\}$, the GRPO objective is:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{o\sim\pi_\theta}\left[\frac{R(o)-\mu_R}{\sigma_R}\log \pi_\theta(o \mid x)\right], \tag{10}$$

where μ_R and σ_R are group-wise statistics over $\{o_1, \ldots, o_G\}$. This encourages stable preference learning and structured reasoning behavior.

5 Experiment

5.1 Experimental Settings

To comprehensively evaluate ClueAegis, we adopt Qwen3.5-9B as the backbone model and conduct all experiments based on this unified configuration. We train the model on ClueAegis-Bench (96K samples) and evaluate on three primary datasets: ClueAegis-Bench, the first skill-oriented dataset; AIGI-Now (Chen et al., 2025a), a dataset built with recent state-of-the-art generators; and GenImage (Zhu et al., 2023), a large-scale dataset with over one million samples. We further include evaluations on six additional datasets in Section 5.5 (training details in *Appendix C*).

Baselines. We compare ClueAegis with a range of state-of-the-art synthetic image detection methods, including CNNSpot(Wang et al., 2020a), FreqNet(Tan et al., 2024a), UnivFD(Ojha et al., 2023), NPR(Tan et al., 2024b), HyperDet(Cao et al., 2024), AIDE(Yan et al., 2024), DDA(Chen et al., 2025b), AIGI-Holmes(Zhou et al., 2025b), REVEAL(Cao et al., 2025), Veritas(Tan et al., 2025) and Ivy-xDetector(Jiang et al., 2025) as representative baselines.

Evaluation Metrics. Following prior work in synthetic image detection, we report Accuracy (ACC) as the primary metric in the main paper, with F1-score results provided in Appendix D. For ClueAegis and LLM-based baselines, we map textual outputs (i.e., *Real* or *Fake*) into binary predictions for evaluation. For non-LLM methods, we follow their official decision thresholds. We do not report AP since several reasoning-based methods produce textual outputs without calibrated confidence scores, making AP not directly applicable under a unified setting. Additional results, including few-shot performance evaluation and qualitative results, are provided in *Appendix E and H*.

5.2 Comparison to State-of-the-Art Detectors

Evaluation on ClueAegis-Bench. Table 1 presents a comparison between ClueAegis and existing synthetic image detection methods on *ClueAegis-Bench*. ClueAegis achieves the best overall performance, surpassing the previous state-of-the-art method by **6.99%** in accuracy. The results demonstrate that ClueAegis effectively learns skill-driven forensic representations, enabling more accurate authenticity reasoning.

Generalization to AIGI-Now. As shown in Table 2, ClueAegis demonstrates strong generalization to recently advanced generative models. It surpasses the previous state-of-the-art method by

Detector	FLUX-dev		FLUX-kera		FLUX-kontext		FLUX-pro		gpt4o		jimeng		keling		minimax		Nano		Avg.
	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	
Conventional Binary Detection Methods																			
CNNSpot	76.90	74.85	78.30	59.05	62.70	57.35	64.80	63.95	74.90	61.00	68.75	60.65	75.00	72.35	72.40	57.65	76.45	70.75	68.21
FreqNet	72.80	52.15	72.50	51.35	60.15	53.45	49.05	50.30	74.80	55.50	56.30	52.35	73.85	53.65	68.55	52.25	76.20	50.05	59.74
UnivFD	85.20	89.75	77.70	85.40	62.05	73.10	61.90	87.55	86.25	91.85	70.25	91.75	85.05	90.60	68.10	80.75	83.95	92.40	81.31
NPR	54.85	48.05	55.85	46.15	52.80	44.75	48.45	48.15	54.20	49.40	51.10	43.95	54.20	63.00	53.75	44.50	55.20	55.15	51.31
AIDE	87.70	72.50	86.85	61.90	76.20	61.50	71.25	70.55	86.25	73.40	71.15	71.30	87.10	79.30	81.60	63.55	88.95	86.55	76.53
DDA	78.30	89.70	78.45	65.90	73.90	64.50	63.10	86.20	78.00	86.15	60.85	81.45	78.35	87.30	75.95	65.85	79.45	90.05	76.86
MLLM-based Detection Methods																			
AIGI-Holmes	98.70	85.30	90.60	74.80	88.40	74.15	**99.00**	80.45	**99.65**	79.60	**99.60**	85.00	97.50	87.20	88.05	65.90	**96.55**	**96.15**	88.14
REVEAL	92.50	91.00	94.00	84.65	87.65	88.00	79.95	74.55	92.35	87.40	91.15	90.00	93.65	88.80	66.95	75.65	80.45	83.35	85.67
Veritas	84.70	81.35	68.05	75.40	69.20	73.85	60.75	77.40	69.15	80.70	68.10	81.40	84.70	82.80	60.25	73.80	81.65	83.20	75.36
Ivy-xDetector	95.40	88.35	94.80	86.85	85.45	81.30	90.25	87.45	95.45	87.75	95.65	87.90	95.55	87.80	92.50	86.80	95.30	87.50	90.11
ClueAegis	**98.85**	**99.75**	**98.70**	**99.45**	**96.20**	**98.90**	90.60	**99.20**	98.75	**99.50**	92.05	**99.15**	**98.35**	**99.15**	**97.90**	**99.45**	83.95	92.40	**97.99**

Table 2: **Generalization on AIGI-Now.** ClueAegis achieves the best overall performance in terms of accuracy, demonstrating superior generalization capability across advanced generative models.

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot	75.46	70.31	69.24	60.10	68.39	69.97	49.83	45.45	63.59
FreqNet	66.70	62.45	60.90	41.08	58.81	61.14	53.07	47.97	56.52
UnivFD	76.25	75.36	73.34	58.91	63.46	73.20	63.74	57.52	67.72
NPR	60.79	61.26	58.40	61.73	59.42	60.29	57.96	50.33	58.77
HyperDet	45.61	56.53	53.10	62.12	37.44	61.60	60.43	33.94	51.35
AIDE	82.23	79.34	79.44	73.94	82.99	78.66	74.49	66.09	77.15
DDA	80.64	79.73	78.74	75.00	76.34	78.43	79.23	79.62	78.47
AIGI-Holmes	86.10	93.17	91.22	84.32	72.53	92.10	89.77	91.00	87.53
REVEAL	93.75	**97.81**	**97.19**	95.00	86.88	96.25	95.94	96.88	94.96
Veritas	66.06	79.73	81.11	90.11	82.46	82.85	83.84	85.63	81.47
Ivy-xDetector	95.00	85.54	86.72	94.44	94.81	94.38	94.68	93.52	92.39
ClueAegis	**97.60**	92.79	91.76	**96.38**	**96.56**	**96.28**	**96.35**	**97.30**	**95.63**

Table 3: **Performance on GenImage.** ClueAegis achieves the best overall performance on the GenImage benchmark, surpassing existing methods by **0.67%** in accuracy and demonstrating strong generalization capability on conventional synthetic image detection settings.

7.88% in accuracy. The improvement is attributed to the heuristic-to-reasoning paradigm, which enables more robust forensic reasoning under unseen generative artifacts.

Performance on GenImage. As shown in Table 3, we evaluate ClueAegis on the GenImage benchmark under standard synthetic image detection settings. ClueAegis consistently achieves the best overall performance compared with existing methods. The results indicate strong and stable performance across diverse synthetic image sources.

5.3 Ablation Studies

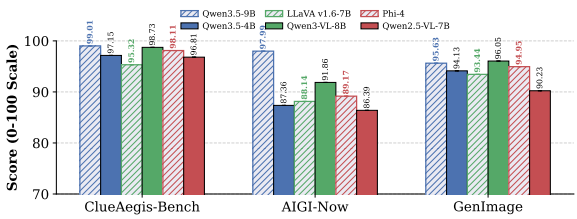


Figure 4: Performance with different backbones.

Effect of Different Backbone Models. As shown in Figure 4, ClueAegis consistently benefits from stronger backbone models across the Qwen series, Phi-4 (Abouelenin et al., 2025), and LLaVA-1.6 (Liu et al., 2023). Overall, larger and more capable vision-language models achieve better detection performance, especially on challenging unseen benchmarks such as AIGI-Now. This suggests that ClueAegis relies on strong multimodal perception and effective skill activation to support forensic reasoning.

Effect of Forensic Skill Numbers. As shown in Table 4, we evaluate the impact of different numbers of forensic skills by reporting the average performance across ClueAegis-Bench, AIGI-Now, and GenImage. Experimental results show that increasing the number of forensic skills improves detection performance, demonstrating that richer skill-aware cognitive routing enables the model to capture more diverse and heterogeneous forgery patterns across synthetic image distributions.

Effect of Training Strategy. As shown in Table 5, we report the average accuracy across ClueAegis-Bench, AIGI-Now, and GenImage to analyze the effectiveness of different training strate-

Skill Configuration	AEGIS-Bench	REVEAL-Bench	GenImage
None (0 Skills)	87.31	90.94	89.79
Compact (4 Skills)	91.53	94.73	93.24
Advanced (8 Skills)	95.68	95.74	94.28
Full (12 Skills)	99.01	97.99	95.63

Table 4: Effect of cognitive skill scale on detection performance across different benchmarks.

Base Model	Skill	SFT	GRPO Alignment	ACC (%)
✓	✗	✗	✗	57.82
✓	✓	✗	✗	59.97
✓	✗	✓	✗	89.84
✓	✓	✓	✗	96.40
✓	✓	✓	✓	97.54

Table 5: Ablation of training components.

gies. Specifically, we compare the performance of the base model, skill-guided inference, supervised fine-tuning (SFT), and GRPO-based cognitive alignment. Experimental results show that enabling the model to first perform cognitive perception and subsequently activate skill-specific forensic reasoning significantly improves synthetic image detection performance. Furthermore, SFT and GRPO alignment further enhance reasoning consistency and generalization capability across diverse synthetic image distributions.

5.4 Robustness Evaluation

We evaluate the robustness of ClueAegis on ClueAegis-Bench under common image perturbations, including JPEG compression, Gaussian blur, and resizing, which simulate real-world distribution shifts that often degrade synthetic image detection performance. As shown in Table 6, ClueAegis consistently outperforms all baselines across corruption types and severity levels, demonstrating strong stability under distribution shift. In contrast, existing methods suffer noticeable performance degradation, particularly under heavy blur and resizing. These results demonstrate that the skill-conditioned framework is robust to input corruption, as skill-level reasoning operates on abstract forensic cues that remain more invariant to low-level perturbations, rather than relying on fragile pixel-level patterns.

5.5 Evaluation on Additional Datasets

We further evaluate ClueAegis on six additional datasets (ForenSynths (Wang et al., 2020a), AIGCDetectBenchmark (Zhong et al., 2023), Chameleon (Yan et al., 2024), FakeBench (Li et al., 2025b), LOKI (Ye et al., 2025), and CommunityAI (Li et al., 2026)) and compare it with several state-of-the-art methods. As shown in Fig. 5, ClueAegis achieves the best or highly competitive performance across all datasets, obtaining the top results on most benchmarks while remaining comparable to the strongest baseline on ForenSynths. These results demonstrate its strong generalization ability under diverse data distributions.

Method	JPEG Compression		Gaussian Blur		Resize
	QF=85	QF=70	$\sigma = 1.0$	$\sigma = 2.0$	$\times 0.5$
UnivFD	85.19	83.62	86.41	84.75	86.13
DDA	90.95	88.22	91.54	85.24	92.98
AIGI-Holmes	79.64	77.01	83.36	78.12	80.69
Veritas	75.86	74.92	76.71	74.37	62.13
ClueAegis	95.43	92.84	92.84	86.25	94.87

Table 6: Robustness comparison in terms of ACC under common image perturbations.

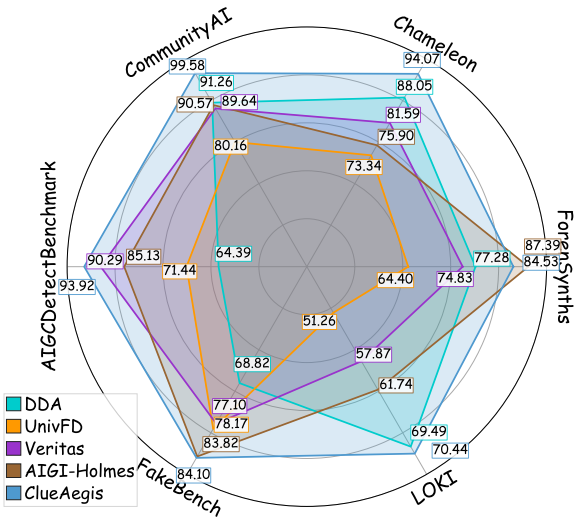


Figure 5: Results on additional benchmarks.

6 Conclusion

In this paper, we revisit synthetic image detection from a heuristic-to-reasoning perspective and propose **ClueAegis**, a skill-driven forensic framework that transforms end-to-end detection into structured reasoning. By decomposing authenticity analysis into forensic skills, **ClueAegis** enables adaptive skill routing and evidence verification across diverse forgeries. We further introduce **ClueAegis-Bench**, a large-scale benchmark with explicit skill annotations for structured detection. Experiments demonstrate SOTA performance with strong generalization across diverse synthetic distributions. This work promotes skill-aware reasoning for trustworthy synthetic media detection.

Limitations

Despite its effectiveness, ClueAegis has several limitations. First, the framework relies on a predefined set of forensic skills, which may not fully cover emerging or unseen manipulation patterns, limiting adaptability to novel generation techniques. Second, the performance of ClueAegis is partially dependent on the capability of the underlying vision-language backbone; weaker models may struggle to reliably activate and ground the correct forensic skills, leading to degraded reasoning quality. Third, the current skill annotations in ClueAegis-Bench are manually designed and may introduce bias or ambiguity in borderline cases where multiple forensic cues coexist. Finally, the structured reasoning pipeline introduces additional computational overhead compared to direct binary classification, which may affect scalability in real-world low-latency scenarios.

Ethics Statement

This work aims to improve the reliability and robustness of synthetic image detection for trustworthy AI-generated media analysis. ClueAegis and ClueAegis-Bench are designed to support forensic reasoning research and help mitigate risks related to misinformation and manipulated visual content. However, synthetic image detection systems may still produce false predictions, especially under unseen generation distributions or ambiguous forensic cues. In addition, the predefined forensic skill annotations may introduce subjective bias in certain cases. We therefore encourage future work on more adaptive, fair, and robust reasoning mechanisms for synthetic media forensics. All datasets used in this work are collected from publicly available sources, and no private or personally sensitive information is intentionally introduced.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62502442, U24A20326), Zhejiang Provincial Natural Science Foundation of China (No. LQN26F020009), and the Science-Tech Innovation Community of the Yangtze River Delta (No. 2023CSJZN0301).

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, and 1 others. 2025. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*.

Huangsen Cao, Qin Mei, Zhiheng Li, Yuxi Li, Zhan Meng, Ying Zhang, Chen Li, Zhimeng Zhang, Xin Ding, Yongwei Wang, and 1 others. 2025. Reveal: Reasoning-enhanced forensic evidence analysis for explainable ai-generated image detection. *arXiv preprint arXiv:2511.23158*.

Huangsen Cao, Yongwei Wang, Yinfeng Liu, Sixian Zheng, Kangtao Lv, Zhimeng Zhang, Bo Zhang, Xin Ding, and Fei Wu. 2024. Hyperdet: Generalizable detection of synthesized images by generating and merging a mixture of hyper loras. *arXiv preprint arXiv:2410.06044*.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. 2020. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pages 103–120. Springer.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. 2025a. Task-model alignment: A simple path to generalizable ai-generated image detection. *arXiv preprint arXiv:2512.06746*.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and 1 others. 2025b. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang,

Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pages 23247–23275.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Chenyang Si, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, and 1 others. 2025. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. *arXiv preprint arXiv:2506.00979*.

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

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

Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Shuchang Lyu, Baoyuan Wu, and Guangliang Cheng. 2025a. Raidx: A retrieval-augmented generation and grpo reinforcement learning framework for explainable deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11746–11755.

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

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. 2026. Is artificial intelligence generated image detection a solved problem? *Advances in Neural Information Processing Systems*, 38.

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

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780.

Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in neural information processing systems*, 36:25435–25447.

Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. 2026. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*.

Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallendael, Peter Lambert, and Symeon Papadopoulos. 2024. Tgif: Text-guided inpainting forgery dataset. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Huaxiao Mo, Bolin Chen, and Weiqi Luo. 2018. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM workshop on information hiding and multimedia security*, pages 43–47.

Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2307–2311. IEEE.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24480–24489.

Rémi Pautrat, Daniel Barath, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. 2023. Deeplsd: Line segment detection and refinement with deep image gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17327–17336.

Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. 2024. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28140–28149.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

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

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024b. 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*, pages 28130–28139.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. 2025. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*.

Alex Jinpeng Wang, Dongxing Mao, Weiming Han, Jiawei Zhang, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, and 1 others. Textatlas5m: A large-scale dataset for long and structured text image generation.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020a. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704.

Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. 2020b. Instance shadow detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1880–1889.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, and 1 others. 2026. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2025. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pages 31186–31216.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2024. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, and 1 others. 2025. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. In *International Conference on Learning Representations*, volume 2025, pages 70440–70522.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. 2025. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*.

Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. 2025. Towards universal ai-generated image detection by variational information

bottleneck network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23828–23837.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. 2019. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. 2023. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*.

Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. 2025a. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. 2025b. 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*, pages 18746–18758.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782.

A Visualization of ClueAegis-Bench

Figure 6 presents the visualization of ClueAegis-Bench. The benchmark organizes images into skill-driven forensic subspaces according to distinct cognitive reasoning cues, covering diverse synthetic image distributions and visual contents. Each subspace corresponds to a specific forensic skill and is associated with structured reasoning annotations, enabling skill-aware evaluation beyond conventional real-versus-fake classification. Specifically, *Lighting Consistency* focuses on illumination-related artifacts and abnormal lighting patterns; *Shadow Consistency* examines the coherence between shadows and corresponding objects; *Physical Consistency* analyzes violations of physical laws and spatial plausibility; *Common Sense Consistency* targets semantically unreasonable or commonsense-defying content; *Functional Consistency* evaluates incorrect functional structures or tool usage; *OCR Consistency* analyzes textual content and typography within images; *Human Anatomy* and *Animal Anatomy* focus on structural abnormalities in humans and animals, respectively; *Region Analysis* investigates local visual artifacts and regional inconsistencies; *Frequency Consistency* performs spectrum-domain forensic analysis; *Pixel Consistency* examines low-level pixel distributions and texture continuity; and *Transformation Consistency* focuses on inconsistencies introduced by post-processing operations such as rotation and color transformation.

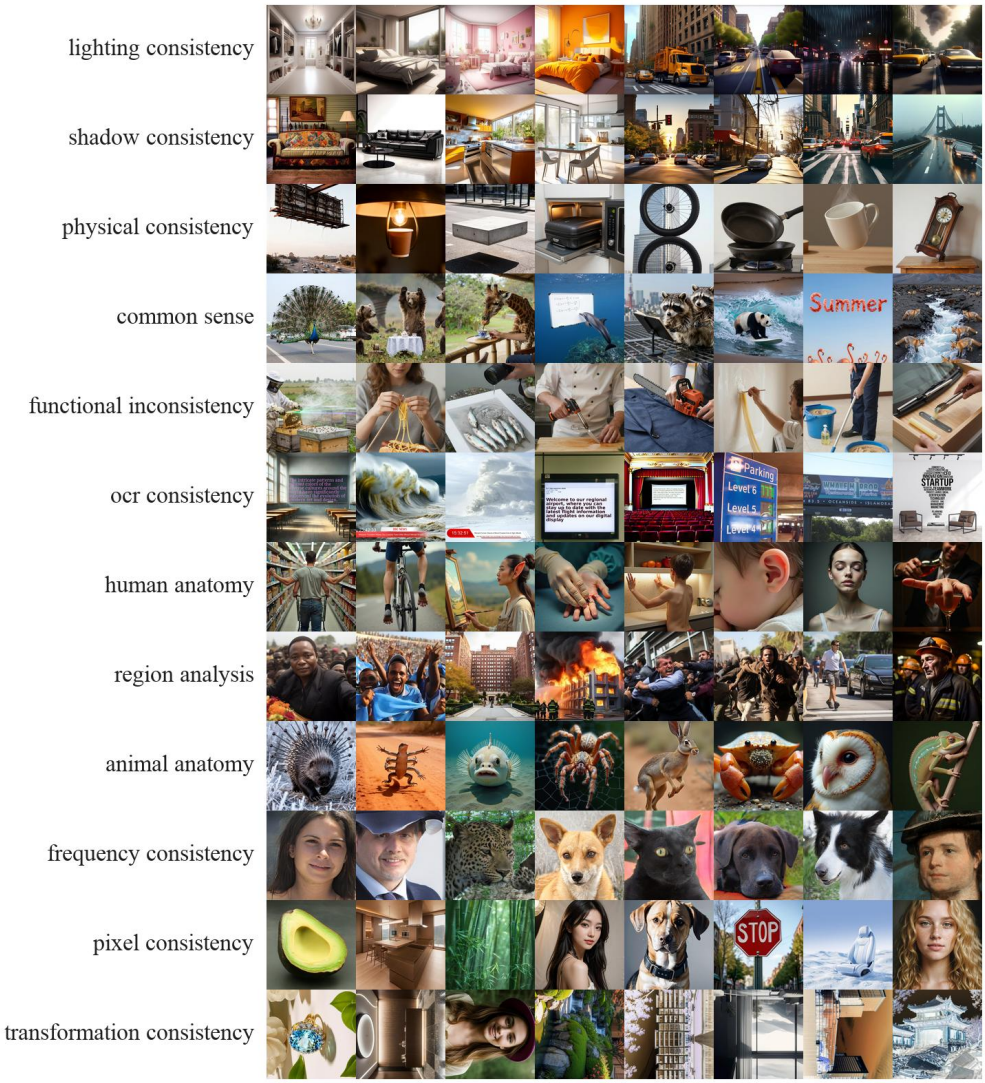


Figure 6: **Visualization of ClueAegis-Bench.** Representative image samples from different skill-driven forensic subspaces with diverse visual contents and synthetic distributions.

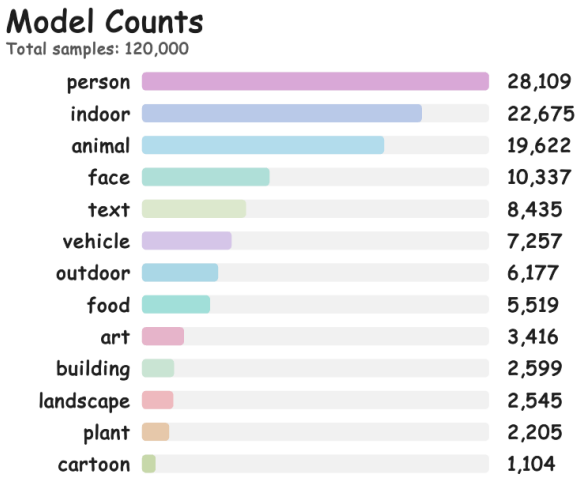


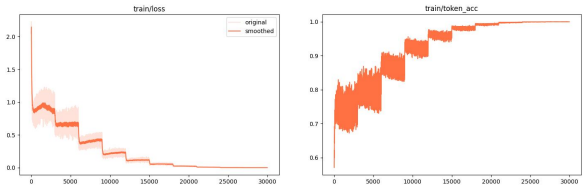
Figure 7: Performance across different visual content types.

B Details of ClueAegis-Bench

As illustrated in Figure 7, we present the distribution of ClueAegis-Bench across different image content categories. In this section, we further describe the detailed annotation pipeline of the benchmark construction process. During data annotation, we first collected a large-scale pool of real and synthetic images from diverse sources. We then utilized the skill definitions provided in Appendix G together with Qwen3-VL-235B to perform skill-oriented annotation. For each candidate image, the model was prompted to determine whether the image could be reliably identified through a specific forensic skill. Only images that could be correctly and consistently recognized under a single corresponding forensic skill were retained. The final benchmark therefore consists of paired `(image, skill)` annotations, where each image is explicitly associated with its dominant forensic reasoning cue. This annotation strategy enables the benchmark to establish a structured mapping between visual evidence and forensic reasoning skills, facilitating skill-aware synthetic image detection and interpretable reasoning evaluation beyond conventional binary classification settings.

C Training Details

In this section, we present the training details of our model, including the configurations for supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO). ClueAegis-Bench contains 120K samples in total, which are split into 96K training samples and 24K held-out test samples. To ensure a fair comparison, all binary classification baseline models are trained on the same 96K training split. For large multimodal LLM-based methods, we follow their standard evaluation protocols without additional training on ClueAegis-Bench, due to their reliance on proprietary training paradigms or pretraining constraints.



(a) Train loss (b) Train token accuracy

Figure 8: Training dynamics of our model. (a) Training loss. (b) Token-level accuracy.

C.1 SFT Training

We fine-tune the model using full-parameter supervised fine-tuning (SFT) on 32 NVIDIA H20 GPUs (96GB memory each) in a distributed multi-node setting. We adopt the Qwen3.5-9B model as the backbone and train with bfloat16 precision. The model is optimized using a learning rate of 1×10^{-4} with a batch size of 32 (via gradient accumulation). We train for 10 epochs with a maximum sequence length of 7168 tokens. The warmup ratio is set to 0.05, and we update all linear modules during training. DeepSpeed ZeRO-2 is employed to reduce memory consumption and enable efficient large-scale training. Figure 8 illustrates the training dynamics, including the convergence behavior of the training loss and token-level accuracy.

C.2 GRPO Training

We further fine-tune the model using Group Relative Policy Optimization (GRPO) on top of the supervised checkpoint. Training is conducted on 8 NVIDIA H20 GPUs (96GB memory each) with bfloat16 precision. We adopt a two-turn interaction setting with a maximum sequence length of 10240 tokens. The model is optimized using a cosine learning rate schedule with an initial learning rate of 1×10^{-6} and trained for 1 epoch. The effective batch size is 32 via gradient accumulation. We employ a group-based reward strategy with multiple signals, including skill consistency, reasoning validity, and answer correctness, to guide policy optimization, where all reward weights are set to 1 without additional hyperparameter tuning. DeepSpeed ZeRO-2 is used for memory-efficient

training.

D F1-score Evaluation Results

In the main paper, we report accuracy (ACC) as the primary metric under a unified real/fake decision rule. To provide a complementary view, Appendix D reports F1 scores calculated from the same binary predictions. For each evaluation slice, F1 is calculated as the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{11}$$

We calculate F1 separately for each forensic skill on ClueAegis-Bench, each generator–cue pair on AIGI-Now, and each generator family on GenImage. The average column follows the aggregation protocol of the corresponding benchmark. These results provide a class-sensitive view of detector performance, especially on slices where class imbalance or uneven difficulty can make accuracy less informative.

D.1 F1-score Evaluation on ClueAegis-Bench

Table 7 mirrors Table 1 by replacing accuracy with F1 on the twelve skill-specific test subsets. ClueAegis achieves the strongest average F1 score (**99.00%**) and ranks first on every skill, indicating consistently strong class-sensitive performance across diverse forensic dimensions. Among prior methods, DDA obtains the highest average F1 score (91.25%), followed by AIDE (89.12%). Several specialized detectors remain competitive on isolated skills, such as CNNSpot on Region and AIGI-Holmes on Freq/Pixel, but their performance varies substantially across other skills. Overall, the F1 results support the accuracy-based conclusion that cognition-guided skill routing is most beneficial when single-feature detectors become brittle.

D.2 F1-score Evaluation on AIGI-Now

Table 8 reports F1 scores on AIGI-Now across recent generators and two cue types, *pix* and *sem*. ClueAegis achieves the best average F1 score (**97.93%**), outperforming the strongest prior baseline, Ivy-xDetector (91.51%), by 6.42% points. The improvement is especially consistent on semantic-cue subsets, where ClueAegis obtains near-saturated F1 scores across all generators. These results suggest that the proposed heuristic-to-reasoning pipeline generalizes beyond low-level artifacts and remains effective on challenging semantic manipulations.

D.3 F1-score Evaluation on GenImage

Table 9 presents F1 scores on the GenImage benchmark. ClueAegis achieves the best mean F1 score (**93.92%**), exceeding Ivy-xDetector (92.35%) by 1.57% points and obtaining the top score on seven of the eight generator families. The only exception is Wukong, where Ivy-xDetector performs best while ClueAegis remains second. This trend shows that ClueAegis retains strong F1 performance under conventional synthetic-image detection settings, including generator families that differ substantially from the skill-oriented training distribution.

E Few-Shot Performance Evaluation

To further examine data efficiency, we evaluate all methods under a few-shot setting on ClueAegis-Bench using the same skill-wise accuracy protocol as the main evaluation. As shown in Table 10, ClueAegis achieves the best average accuracy (**97.60%**) and ranks first across all twelve forensic skills. Compared with the strongest baseline, DDA (92.42%), ClueAegis improves the average accuracy by 5.18% points. The gains are especially pronounced on OCR, Region, Pixel, and Trans, where reliable detection requires reasoning over localized evidence or structural transformations rather than relying only on global visual artifacts. These results indicate that the proposed heuristic-to-reasoning framework remains effective when only limited supervision is available, suggesting stronger data efficiency and skill-level generalization than conventional detectors.

F Human Expert Verification

To further validate the reliability of the forensic skill annotations in ClueAegis-Bench, we conducted an independent human verification study on 120 randomly sampled images. Three independent experts with research experience in synthetic image detection and image forensics participated in the evaluation. The experts independently examined each image and judged whether the assigned forensic skill was supported by the visual evidence. No expert had access to the judgments of the other experts during annotation. When disagreements occurred, the final annotation was determined by majority voting. As shown in Table 11, the three experts achieved an average agreement of 89.45%,

Table 7: **F1-score Evaluation on ClueAegis-Bench.** F1 is computed for each skill-specific test subset with *Fake* as the positive class.

Method	Forensic Consistency Metrics												Avg.
	Light	Shadow	Phys	CS	Func	OCR	Human	Region	Animal	Freq	Pixel	Trans	
CNNSpot	94.35	93.29	91.70	92.39	91.80	92.20	92.72	87.56	92.50	65.35	91.33	82.07	88.94
FreqNet	84.43	84.75	89.79	93.68	92.50	87.06	92.37	32.73	90.49	86.55	88.75	75.03	83.18
UnivFD	93.49	91.37	86.95	94.28	90.39	90.39	91.68	71.73	92.68	81.67	82.17	75.56	86.86
NPR	87.21	82.42	88.37	89.16	88.90	85.21	88.67	82.10	88.53	89.12	86.04	84.68	86.70
HyperDet	1.19	0.40	84.85	88.65	84.49	1.78	89.67	0.80	89.21	84.68	75.39	54.47	54.63
AIDE	98.89	90.62	94.20	94.50	94.20	93.18	94.50	45.92	94.25	93.50	92.33	83.38	89.12
DDA	94.50	93.70	94.45	94.65	94.25	93.10	94.50	63.30	94.50	94.35	93.58	90.18	91.25
AIGI-Holmes	67.45	67.14	86.18	89.90	80.49	90.73	90.75	82.72	80.68	98.96	97.76	84.69	84.79
Veritas	87.03	79.12	71.69	79.20	70.85	51.57	85.68	68.39	77.02	76.55	70.64	64.32	73.50
Ivy-xDetector	66.96	67.07	93.31	95.00	91.73	63.52	94.93	81.68	95.16	76.65	90.33	73.16	82.46
ClueAegis	**99.80**	**99.70**	**99.45**	**99.70**	**99.75**	**99.85**	**99.70**	**95.49**	**99.75**	**99.95**	**99.50**	**95.35**	**99.00**

Table 8: **F1-score Evaluation on AIGI-Now.** F1 is computed for each generator–cue pair with *Fake* as the positive class.

Detector	FLUX-dev		FLUX-kera		FLUX-kontext		FLUX-pro		gpt4o		jimeng		keling		minimax		Nano		Avg.
	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	
Conventional Binary Detection Methods																			
CNNSpot	81.03	66.49	82.48	30.77	65.21	26.27	69.04	43.72	79.68	36.17	73.44	35.33	79.76	61.99	77.21	26.79	80.59	58.72	59.71
FreqNet	77.39	12.60	76.75	10.32	61.96	16.05	48.35	7.28	79.56	23.67	58.73	14.83	78.71	18.47	73.36	15.11	80.43	5.84	42.19
UnivFD	89.80	94.53	80.30	89.53	54.58	72.35	55.21	92.07	91.06	96.79	69.95	96.70	89.78	95.45	66.50	83.72	88.60	97.39	83.57
NPR	68.83	42.44	69.37	37.05	66.71	36.09	63.30	43.92	68.59	45.88	65.83	37.48	68.59	65.36	68.25	37.99	69.04	51.59	55.91
AIDE	88.86	62.28	87.52	39.14	75.71	37.60	69.85	58.61	87.54	64.10	69.71	59.97	88.34	74.03	82.51	43.27	89.76	84.60	70.19
DDA	82.07	88.74	82.21	49.33	77.24	46.13	65.71	84.32	81.97	84.27	62.84	77.66	82.13	85.81	79.87	49.96	82.95	89.19	75.13
MLLM-based Detection Methods																			
AIGI-Holmes	98.68	82.61	89.63	66.62	87.01	65.82	**98.89**	76.09	**99.55**	75.02	**99.75**	82.24	97.34	85.21	86.46	49.55	96.38	90.01	84.83
Veritas	84.13	83.16	60.63	76.21	60.73	74.77	47.49	79.36	63.38	82.52	60.95	83.29	84.39	84.63	45.96	74.71	80.90	85.31	72.92
Ivy-xDetector	95.59	89.51	94.94	88.17	92.23	82.70	93.79	88.66	95.62	88.92	95.80	89.19	95.72	89.03	94.85	88.09	95.51	88.92	91.51
ClueAegis	**98.71**	**99.45**	**98.72**	**99.45**	**96.18**	**99.65**	89.98	**99.21**	98.42	**99.21**	91.65	**99.21**	**98.38**	**99.16**	**97.92**	**99.45**	**98.33**	**99.65**	**97.93**

Table 9: **F1-score Evaluation on GenImage.** F1 is computed for each generator family with *Fake* as the positive class.

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot	77.71	72.79	71.87	59.23	69.55	67.95	42.43	33.84	61.92
FreqNet	65.06	59.35	57.59	17.71	53.08	57.70	42.64	32.88	48.25
UnivFD	75.70	74.95	72.53	49.66	58.25	72.27	58.61	48.64	63.83
NPR	71.30	71.55	69.77	71.90	69.82	70.69	68.58	60.49	69.26
AIDE	82.02	78.70	79.00	71.81	83.42	78.14	72.56	58.79	75.55
DDA	83.34	83.08	82.23	78.29	79.43	82.00	82.74	83.03	81.77
Veritas	50.18	61.87	63.56	61.60	63.54	71.53	69.14	66.33	63.47
Ivy-xDetector	94.25	85.71	87.44	93.47	95.14	**95.48**	94.64	92.64	92.35
ClueAegis	**96.33**	**90.31**	**90.48**	**95.27**	**96.48**	90.59	**96.06**	**95.87**	**93.92**

Table 10: **Few-Shot Performance Evaluation.** Accuracy is reported on ClueAegis-Bench under the few-shot setting.

Method	Forensic Consistency Metrics												Avg.
	Light	Shadow	Phys	CS	Func	OCR	Human	Region	Animal	Freq	Pixel	Trans	
CNNSpot	91.60	92.10	90.10	90.70	91.05	92.30	91.15	87.80	90.50	89.35	85.79	75.15	88.97
FreqNet	72.90	70.50	86.10	91.20	90.65	83.50	92.60	77.95	91.35	79.60	81.73	71.55	82.47
UnivFD	93.50	91.50	87.05	90.55	89.30	85.70	91.70	71.95	90.10	76.50	77.23	70.90	84.67
NPR	85.25	75.30	90.05	92.55	91.80	67.85	92.95	78.00	91.00	90.60	84.69	71.25	84.27
HyperDet	59.15	49.20	91.20	94.10	93.00	58.25	92.25	69.90	90.05	87.85	83.89	75.80	78.72
AIDE	93.30	91.90	91.35	92.65	91.45	88.50	93.15	76.15	92.25	91.15	88.29	77.85	89.00
DDA	92.65	93.90	93.90	94.40	93.90	91.35	94.20	88.10	94.30	93.90	92.40	86.00	92.42
ClueAegis	**98.55**	**99.00**	**97.40**	**98.35**	**97.75**	**99.55**	**96.70**	**94.35**	**96.55**	**98.90**	**98.85**	**95.30**	**97.60**

	Expert 1	Expert 2	Expert 3	Majority Vote	Average
Agreement (%)	89.17	86.67	92.50	**90.00**	**89.45**

Table 11: Human expert verification of forensic skill annotations on 120 randomly sampled images from ClueAegis-Bench.

while the majority-vote agreement reached 90.00%. These results demonstrate that the skill annotations are generally well aligned with expert human judgments.

G Skill Design Details

This appendix reports the skill-conditioned second-round prompt templates used by ClueAegis. To avoid repeating identical instructions, we factor out the shared wrapper: each prompt is applied after the first-round image analysis, requires the model to compare the current evidence with the previous-round original image, and follows the same answer format, where the reasoning process is generated within the `<think>` block, and the final answer outputs `0` if the image is real and `1` if the image is synthetic.

We present the detailed prompt designs of all twelve forensic skills in this appendix. Different forensic skills are associated with different auxiliary analysis modules and external evidence extraction tools. Specifically, *Lighting Consistency* incorporates an external lighting analysis model(Pautrat et al., 2023) for illumination reasoning; *Shadow Consistency* utilizes external object and shadow extraction models(Wang et al., 2020b) to analyze shadow-object coherence; *OCR Consistency* employs an OCR model(Cui et al., 2025) to extract textual information; *Region Analysis* integrates a regional perception model(Kirillov et al., 2023) to capture localized object evidence; *Frequency Consistency* incorporates a frequency-spectrum analysis model(Tan et al., 2024a) for spectral forensic detection; and *Pixel Consistency* leverages a pixel-level forensic model(Tan et al., 2024b) to identify low-level pixel artifacts and local inconsistency patterns.

Skill 1: Lighting Consistency

```
Input tokens: <image><image>
Images:
- Image 1: original image.
- Image 2: lighting-consistency auxiliary
    view from the DeepLSD line-based analysis.

Task: Determine whether the image is real or
    synthetic using lighting consistency
    evidence.

Reasoning checklist:
1. Use Image 1 as the reference for scene
    structure and illumination context.
2. Compare Image 2 with Image 1 for lighting
    direction, intensity, highlights, and
    reflections.
3. Check whether visible shadows and
    geometric cues are consistent with a
    coherent light source.
4. Identify unrealistic lighting patterns,
    abrupt illumination changes, or
    physically implausible light behavior.
5. Decide whether lighting inconsistencies
    provide evidence of synthetic generation.
```

Skill 2: Shadow Consistency

```
Input tokens: <image><image><image>
Images:
- Image 1: original image.
- Image 2: instance/object mask from the
    shadow detector.
- Image 3: predicted shadow mask from the
    same detector.

Task: Determine whether the image is real or
    synthetic using shadow consistency
    evidence.

Reasoning checklist:
1. Align object regions in Image 2 and shadow
    regions in Image 3 with the original
    scene.
2. Check whether predicted shadows match
    visible shadows in direction, shape,
    softness, and attachment.
3. Use the object mask to judge whether
    shadow regions correspond to plausible
    casters.
4. Identify missing, detached, or
    directionally inconsistent shadows.
5. Decide whether shadow-related
    inconsistencies indicate synthetic
    generation.
```

Skill 3: Physical Consistency

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.

Task: Determine whether the image is real or
    synthetic using physical consistency
    reasoning.

Reasoning checklist:
1. Recall the original image from the first
    round.
2. Evaluate geometric structure and spatial
    layout consistency.
3. Check lighting behavior and shadow
    plausibility under a single physical
    light source.
4. Analyze object scale, perspective
    coherence, and depth relationships.
5. Identify violations of real-world physical
    constraints.
6. Decide whether the scene obeys real-world
    physical laws.
```

Skill 4: Common-Sense Consistency

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.

Task: Determine whether the image is real or
    synthetic using common-sense reasoning.

Reasoning checklist:
1. Recall the original image from the first
    round.
2. Evaluate whether the scene, objects, and
    interactions are plausible in real-world
    contexts.
3. Check whether object presence,
    relationships, and functions match
    everyday experience.
4. Identify semantic contradictions such as
    impossible object usage, unrealistic
    combinations, or illogical arrangements.
5. Verify that the scene remains semantically
    consistent across rounds.
6. Decide whether the overall scene makes
    sense under normal human understanding.
```

Skill 5: Functional Consistency

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.

Task: Determine whether the image is real or
    synthetic using functional consistency
    reasoning.

Reasoning checklist:
1. Recall the original image from the first
    round.
2. Evaluate whether objects function in
    physically and semantically valid ways.
3. Check whether object parts are used
    correctly according to real-world
    functions.
4. Identify incorrect usage, non-functional
    structures, impossible operations, or
    broken interactions.
5. Verify that functional relationships
    remain consistent across rounds.
6. Decide whether functional inconsistencies
    indicate synthetic generation.
```

Skill 6: OCR Consistency

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.
Auxiliary text:
<text>
__OCR_TEXT__
</text>

Task: Determine whether the image is real or
    synthetic using OCR consistency reasoning.

Reasoning checklist:
1. Compare OCR-extracted text with visible
    text regions in the image.
2. Check for incorrect characters, missing
    text, distorted words, or mismatched
    semantics.
3. Evaluate whether text layout, font,
    alignment, and appearance are visually
    plausible.
4. Assess whether the extracted text is
    contextually consistent with signs,
    labels, or captions in the scene.
5. Verify that text content and visual
    grounding remain consistent across rounds.
6. Decide whether OCR-image inconsistencies
    indicate synthetic generation.
```

Skill 7: Human Anatomy

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.

Task: Determine whether the image is real or
    synthetic using human anatomy consistency
    reasoning.

Reasoning checklist:
1. Recall the original image from the first
    round.
2. Evaluate whether human body structure is
    anatomically plausible.
3. Check proportions of limbs, hands,
    fingers, head, and torso.
4. Identify extra or missing fingers,
    distorted joints, unnatural poses, or
    inconsistent limb connections.
5. Assess facial structure, symmetry, and
    alignment of eyes, nose, and mouth.
6. Decide whether the human anatomy follows
    real-world biological structure.
```

Skill 8: Region Analysis

```
Input tokens: <image><image>
Images:
- Image 1: original image.
- Image 2: local region cropped from the
    original image by the SAM-based crop tool.

Task: Determine whether the image is real or
    synthetic using region-level forensic
    analysis.

Reasoning checklist:
1. Locate where Image 2 corresponds within
    Image 1.
2. Understand the semantic and structural
    context of the selected local region.
3. Examine local texture, edges, color
    transitions, and noise patterns.
4. Identify blurring, over-smoothing,
    repetition, distortion, or unnatural
    boundaries.
5. Compare local details with the global
    context for texture, lighting, and
    structure consistency.
6. Decide whether local-global
    inconsistencies indicate synthetic
    generation.
```

Skill 9: Animal Anatomy

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.

Task: Determine whether the image is real or
    synthetic using animal anatomy
    consistency reasoning.

Reasoning checklist:
1. Recall the original image from the first
    round.
2. Evaluate whether the animal body structure
    is anatomically plausible.
3. Check proportions of head, limbs, torso,
    tail, wings, or other visible body parts.
4. Identify extra or missing limbs, distorted
    joints, unnatural body connections, or
    inconsistent symmetry.
5. Assess whether posture, pose, and
    species-specific traits are realistic.
6. Decide whether the animal anatomy follows
    real-world biological structure.
```

Skill 10: Frequency Consistency

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.
Auxiliary result:
- Frequency detection result: __CALL_ANSWER__
     (external FreqNet result: fake or real)

Task: Determine whether the image is real or
    synthetic using frequency consistency
    reasoning.

Reasoning checklist:
1. Analyze high-frequency details, noise
    patterns, and texture distribution.
2. Identify over-smoothing, repetitive
    patterns, unnatural frequency
    attenuation, or abnormal high-frequency
    noise.
3. Treat the external frequency detector
    output as supporting evidence, not as the
    sole decision.
4. Check whether visual evidence is
    consistent with the frequency-domain
    assessment.
5. Verify that image evidence remains
    consistent with the previous-round
    original image.
6. Decide whether the combined evidence
    indicates synthetic generation.
```

Skill 11: Pixel Consistency

```
Input tokens: <image>
Images:
- Image 1: same original image as in the
    previous round.
Auxiliary result:
- Pixel detection result: __CALL_ANSWER__  (0
    = real, 1 = synthetic according to the
    NPR-based model)

Task: Determine whether the image is real or
    synthetic using pixel consistency
    reasoning.

Reasoning checklist:
1. Analyze pixel-level noise distribution,
    local texture continuity, and compression
    artifacts.
2. Identify inconsistent noise patterns,
    unnatural smooth regions, pixel
    discontinuities, or abnormal local
    structures.
3. Check whether pixel statistics align with
    natural image formation processes.
4. Treat the external pixel detector result
    as supporting evidence, not as the sole
    decision.
5. Verify that visual evidence remains
    consistent with the previous-round
    original image.
6. Decide whether the combined evidence
    indicates synthetic generation.
```

Skill 12: Transformation Consistency

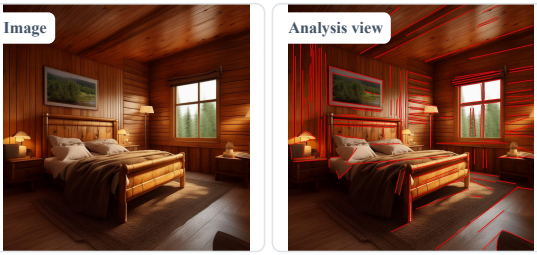
```
Input tokens: <image><image>
Images:
- Image 1: same input image as in the
    previous round, interpreted as the
    transformed view described by
    transform_params.
- Image 2: reconstructed image produced by
    inverting transform_params on Image 1.

Task: Determine whether the image is real or
    synthetic using transformation
    consistency reasoning.

Reasoning checklist:
1. Recall the original image and the
    first-round transform_params.
2. Compare Image 1 with the original image
    for structural, semantic, and visual
    consistency under the reported
    transformation.
3. Compare Image 2 with the original image to
    assess whether reconstruction faithfully
    recovers the original content.
4. Identify missing details, structural
    deviations, information loss, or
    unnatural artifacts introduced by
    transform-recover processing.
5. Analyze whether the transformation and
    reconstruction behave consistently with
    real image processing.
6. Decide whether cross-transformation
    inconsistencies indicate synthetic
    generation.
```

Skill 1: Lighting Consistency

(example1) FAKE



HEURISTIC

<think>The image depicts an interior room scene with a bed, wooden flooring, a window with visible trees outside, two lampshades on bedside tables, and a painting on the wall. The layout includes furniture arranged symmetrically along the walls, with bedding and pillows visible on the bed. The window is centered on the far wall, and the painting hangs above the bed. The overall composition appears clean and orderly. </think><skill_select>Lighting Consistency </skill_select>

REASONING

<think>The lighting in the scene exhibits multiple inconsistencies. The window emits bright, even illumination, yet the shadows on the floor and bed lack directional sharpness and falloff expected from a single natural light source. The lampshades glow with identical intensity despite differing distances and angles relative to the viewer, violating inverse-square lighting behavior. Highlights on wooden surfaces are unnaturally smooth and uniformly distributed, missing the micro-specular reflections typical of real wood finishes. Edge lines in the lighting map (red lines) follow structural boundaries with mathematical precision, especially along wood grain and window frames, suggesting algorithmic edge detection rather than natural photographic detail. The painting and trees visible through the window show unnaturally crisp edges aligned with room geometry, inconsistent with real-world depth-of-field and lens aberrations. These artifacts—perfectly aligned edges, overly smooth lighting, and uniform shadowing—are characteristic of synthetic renderings or AI-generated imagery rather than real-world photography. The image is likely sytheitc.</think> <answer>1</answer>

(example2) REAL

HEURISTIC

<think>The original image shows a bedroom scene containing a bed with pillows and a bedspread, a nightstand with a lamp, a chair, and a painting on the wall. A window is visible on the left side of the room. The layout is typical of a residential bedroom with furniture arranged around the bed. All objects appear in their expected spatial relationships without visible distortion or misplacement.</think><skill_select>Lighting Consistency</skill_select>

REASONING

<think>The lighting in the scene originates consistently from the window on the left, casting soft shadows to the right of objects such as the bed, nightstand, and lamp. Highlights on the bedspread and pillows are diffuse and align with the ambient illumination, showing no unnatural specular reflections. Shadows under the bed and beside the nightstand are soft and proportionate to the light source, indicating natural illumination. Edge analysis confirms clean, geometrically consistent outlines of all objects without misalignment or digital artifacts. The painting on the wall exhibits natural lighting gradients and brushwork consistent with a physical canvas under ambient light. Fabric textures show plausible wrinkles and folds reacting to gravity and lighting. No inconsistencies in shadow direction, highlight placement, or edge integrity are present. Therefore, the scene exhibits realistic lighting behavior consistent with a photograph taken in a real environment.</think> <answer>0</answer>

Figure 9: **Lighting Consistency.** Skill-based forensic reasoning based on illumination consistency analysis. External lighting analysis models are further incorporated to assist the examination of spatial light directions and shadow relationships.

H Qualitative Results

In this section, we present qualitative examples of ClueAegis under all 12 predefined forensic skills. Specifically, we visualize the corresponding reasoning results on both synthesized and real-world images to demonstrate how different forensic skills respond to diverse visual patterns and forensic cues. For each skill category, representative synthetic and real image examples are provided together with the corresponding forensic reasoning process and evidence analysis generated by the model. The qualitative results further illustrate the effectiveness of the proposed skill-driven reasoning framework across heterogeneous image distributions.

Skill 2: Shadow Consistency

(example1) FAKE

Image

Object mask

Shadow map

HEURISTIC

<think>The image shows a living room with a fireplace centered on the back wall, flanked by two symmetrical armchairs. A coffee table is placed in front of the fireplace. A potted plant sits to the right of the right armchair. The floor is covered with a rug that extends under the furniture. The walls are decorated with a brick pattern. The ceiling is not visible. All objects appear clean, with sharp edges and uniform textures. </think><skill_select>Shadow Consistency</skill_select>

REASONING

<think>The image is SYNTHETIC. Shadow direction is uniformly consistent across all objects, with no variation in softness or intensity that would reflect real-world lighting conditions. Shadows are overly diffused and lack contact shadows or ambient occlusion near object bases. Perspective is unnaturally perfect: all furniture aligns symmetrically with no parallax or lens distortion, and object scale remains constant regardless of position, violating depth cues. The fireplace flames are geometrically uniform and lack dynamic variation. Textures (brick, plant, rug) show no micro-imperfections or noise. These physical inconsistencies indicate computer-generated imagery rather than a real photograph.</think> <answer>1</answer>

(example2) REAL

Image

Object mask

Shadow map

HEURISTIC

<think>The image shows a dining room with a wooden table, four chairs, a rug under the table, and large windows in the background. The table has a smooth surface with visible wood grain. The chairs are positioned around the table, two on each side. The rug has a patterned design and lies flat on the floor. The windows are rectangular and span the width of the back wall. Decorative items are visible on the table and along the windowsill. The walls are painted in a neutral tone, and there is no visible ceiling or floor beyond the rug and immediate surroundings.</think><skill_select>Shadow Consistency</skill_select>

REASONING

<think>The image is REAL. Shadow direction is consistent across all objects, with soft, diffuse shadows falling uniformly under the table and chairs, aligned with the ambient light from the windows. Shadow shapes match the contours of the objects casting them, and there is no misalignment or distortion. Perspective cues show consistent vanishing points along the table edges and window frames, with object scale decreasing appropriately with depth. The spatial layout and camera geometry are coherent, with no signs of synthetic rendering artifacts or inconsistent lighting behavior.</think><answer>0</answer>

Figure 10: **Shadow Consistency.** Skill-based forensic reasoning based on shadow consistency analysis. External shadow extraction and object segmentation models are further incorporated to examine the spatial coherence among objects, illumination, and projected shadows.

Skill 3: Physical Consistency

(example1) FAKE

Image

HEURISTIC

<think>The main objects are a cyclist in full riding gear mounted on a road bicycle, captured mid-air above a paved road. The background consists of trees and a clear sky. The cyclist and bike are suspended with no visible contact to the ground. The posture suggests active riding, and there is no visible ramp, motion blur, or propulsion mechanism.</think><skill_select>Physical Consistency</skill_select>

REASONING

<think>The scene is physically implausible. A cyclist and bicycle cannot hover mid-air without a visible source of lift, propulsion, or support. The lack of environmental cues (e.g., ramp, motion blur, landing zone) and the violation of gravity and Newtonian mechanics indicate this is not a real photograph of a physical event, but likely synthetic or digitally manipulated.</think><answer>1</answer>

(example2) REAL

Image

HEURISTIC

<think>The main object in the image is a small bird perched on a rock. The bird is standing upright with its feet gripping the surface of the rock. The background is a blurred, uniform earthy tone, suggesting an outdoor, natural setting. There are no visible liquids or dynamic motions — the bird appears stationary.</think><skill_select>Physical Consistency</skill_select>

REASONING

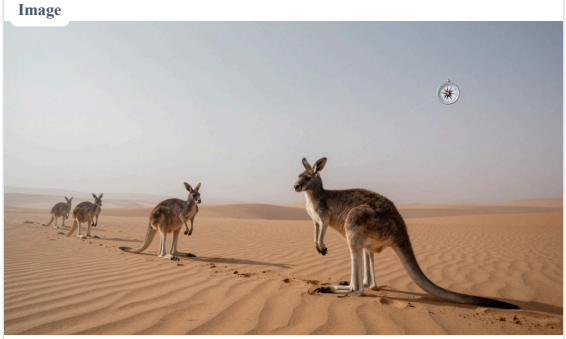
<think>The scene is physically consistent: the bird's posture, grip, and alignment with gravity are anatomically and mechanically plausible. The rock supports the bird's weight without deformation or instability. No unnatural forces, levitation, or fluid dynamics anomalies are present. The image depicts a realistic, naturally occurring moment without violations of physical laws.</think><answer>0</answer>

Figure 11: **Physical Consistency.** Skill-based forensic reasoning from a physical perspective, analyzing whether image content conforms to real-world physical laws and spatial interaction constraints.

Skill 4: Common Sense

(example1) FAKE

Image



HEURISTIC

<think>The caption describes an image featuring kangaroos in a desert environment with a floating compass in the sky connected to one of the kangaroos, noting inconsistencies in lighting, shadows, textures, and ecological plausibility.</think> <skill_select>Common Sense</skill_select>

REASONING

<think>The image is synthetic because kangaroos are not native to vast sand dune deserts like the Sahara, compasses cannot float in the sky connected by lines, lighting and shadows are inconsistent with natural conditions, textures appear unnaturally clean, and the overall composition is surreal — all indicators of digital generation or heavy editing rather than a real photograph.</think><answer>1</answer>

(example2) REAL

Image

HEURISTIC

<think>The image depicts two lionesses in what appears to be a zoo or wildlife enclosure. The foreground lioness is in sharp focus, with natural fur texture, realistic lighting, and shadows consistent with outdoor daylight. The background includes foliage and a wall with faint graffiti, which is slightly out of focus — a natural depth-of-field effect typical of real photography. There are no obvious artifacts, warped anatomy, or unnatural blending that would suggest AI generation. The composition and lighting are consistent with a candid wildlife photograph. The second lioness, partially visible behind the first, also exhibits natural posture and fur detail. No elements appear digitally manipulated or inconsistent with real-world optics.</think> <skill_select>Common Sense</skill_select>

REASONING

<think>The image is consistent with a real photograph, as it exhibits natural lighting, depth of field, anatomical accuracy, and lack of digital artifacts — all indicators of realism rather than synthetic generation.</think><answer>0</answer>

Figure 12: **Common Sense.** Skill-based forensic reasoning based on commonsense consistency analysis, evaluating whether scene content, object relationships, and semantic interactions conform to real-world expectations and everyday human cognition.

Skill 5: Functional Inconsistency

(example1) FAKE

Image

HEURISTIC

<think>A person wearing a welding helmet and gloves holds a welding torch emitting sparks and light, directed at a paper document they are also holding. The paper shows no signs of burning or damage despite direct exposure to the welding arc.</think> <skill_select>Functional Inconsistency</skill_select>

REASONING

<think>The scene is physically and functionally implausible: paper cannot withstand welding temperatures without igniting or vaporizing, and no trained welder would attempt to weld paper. The depicted behavior violates material science, safety norms, and tool functionality, indicating the image is synthetic or artistically staged rather than realistic. </think><answer>1</answer>

(example2) REAL

Image

HEURISTIC

<think>The main action shown is a person leaning over a table, interacting with a laptop computer. The person appears to be using the trackpad or keyboard with one hand while holding a piece of paper or napkin with the other. A beverage can (Pabst Blue Ribbon) sits nearby on the table. The setting appears to be a dimly lit venue, possibly a bar or music venue, with audio equipment visible in the background. </think><skill_select>Functional Inconsistency</skill_select>

REASONING

<think>The scene is realistic: the person's posture, tool usage, and environmental context are all consistent with real-world behavior. No functional inconsistencies are detected — the laptop, beverage can, and paper are appropriately used in a plausible live event setting, and the person's multitasking behavior aligns with human cognition and motor control.</think> <answer>0</answer>

Figure 13: **Functional Inconsistency.** Skill-based forensic reasoning based on functional consistency analysis, evaluating whether objects and scene components conform to realistic functionalities and real-world usage relationships.

Skill 6: OCR Consistency

(example1) FAKE

Image

HEURISTIC

<think>A dramatic volcanic eruption with glowing lava flowing down slopes under a dark, cloudy sky. A banner-like overlay appears across the scene, displaying text elements including a timestamp and two distinct headlines. The banner has sharp edges and uniform alignment, with no visible motion blur or imperfections.</think><skill_select>OCR Consistency</skill_select>

REASONING

<tool>OCR: 18:16:34 Lava Flow from Fuego Volcano Damages Homes in Guatemala New Research Highlights Importance of Early Childhood Education in Success.</tool><think> The visual scene of a volcanic eruption is plausible and consistent with real-world phenomena. However, the OCR-extracted text includes two thematically unrelated headlines fused into a single banner, which is atypical for authentic news broadcasts. Real-time news graphics usually maintain thematic separation or show minor imperfections such as motion blur or variable typography — none of which are present. The clean, stylized design of the text overlay suggests digital composition.</think><answer>1</answer>

(example2) REAL

Image

HEURISTIC

<think>Physical letter cubes arranged on lined paper, with one cube replaced by a die obscuring a letter. Highlighters are visible in the foreground. Natural lighting and shadows are present. Slight misalignment of cubes and depth-of-field blur on foreground objects are evident.</think><skill_select>OCR Consistency</skill_select>

REASONING

<tool>OCR: DO NOT LOOK BACK DO NOT LOOK BACK GO F RWARD GO F RWARD</tool><think> The OCR error ('F RWARD' instead of 'FORWARD') is consistent with visual evidence: the letter 'O' is physically occluded by a die, which OCR systems commonly misinterpret. The scene exhibits real-world characteristics — texture of paper, natural lighting, depth-of-field blur, and slight misalignment — all indicating a real photograph. No synthetic artifacts are present. The inconsistency is explainable by physical occlusion, not generation error. Therefore, the image is likely REAL.

Figure 14: **OCR Consistency.** Skill-based forensic reasoning based on textual consistency analysis. External OCR models are further incorporated to examine recognized text content, typography, and semantic alignment consistency.

Skill 7: Human Anatomy

(example1) FAKE

Image

HEURISTIC

<think>A person's forearm and hand wrapped in white gauze, a gloved hand holding the patient's hand, and a red tubular structure resembling a blood vessel visibly protruding from beneath the skin across the back of the hand, appearing exposed or pulled out from tissue.</think><skill_select>Human Anatomy</skill_select>

REASONING

<think>The depiction is anatomically implausible: human blood vessels are embedded in tissue and cannot be cleanly extracted or appear as free-floating tubes on the skin surface without rupture, bleeding, or surrounding tissue. The scene violates physiological and medical realism, suggesting it is synthetic or artistically rendered rather than real.</think> <answer>1</answer>

(example2) REAL

Image

HEURISTIC

<think>A male figure standing upright outdoors, wearing a white t-shirt, black pants, checkered slip-on shoes, a black baseball cap with a logo, and dark sunglasses, with hands in front pockets, head slightly tilted upward and to the side, in front of a decorative metal gate with greenery and plants in the foreground and background. </think><skill_select>Human Anatomy</skill_select>

REASONING

<think>The figure exhibits anatomically correct proportions, natural posture, and biomechanically plausible positioning consistent with real human physiology and behavior; no distortions or synthetic artifacts are evident, indicating realism.</think> <answer>0</answer>

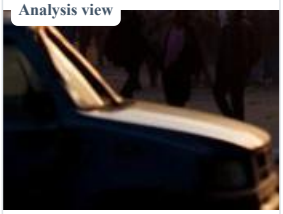
Figure 15: **Human Anatomy.** Skill-based forensic analysis that evaluates whether human body structures, anatomical proportions, and articulation relationships conform to realistic human anatomy principles.

Skill 8: Region Analysis

(example1) FAKE

Image

Analysis view



HEURISTIC

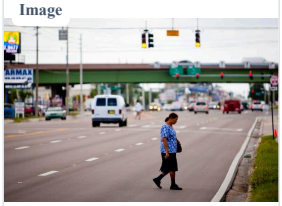
<think>The image shows a chaotic urban scene with a large fire and smoke, a crowd of people many wearing helmets, suggesting a protest or riot. The lighting is dramatic with strong backlighting from the fire, casting silhouettes and creating high contrast. A dark vehicle, likely a police or utility van, is visible in the foreground with reflections and shadows on its hood.</think><skill_select>Region Analysis</skill_select>

REASONING

<think>The cropped region reveals multiple forensic inconsistencies indicating synthetic generation. The vehicle's hood exhibits overly smooth, uniform reflections lacking micro-textural variation expected from real paint under dynamic firelight. Edge boundaries between the vehicle and background figures are unnaturally sharp and partially fused, showing no realistic occlusion or depth-of-field blur. Lighting on the hood is inconsistent with the fire's position — minimal specular highlights despite strong backlighting. Crowd figures behind the vehicle show repetitive helmet shapes and postures, lacking natural variation. Ground debris lacks proper shadows and depth integration, appearing pasted. These artifacts — in texture, lighting, edge coherence, and crowd realism — are characteristic of AI-generated imagery rather than a real photograph.</think>
<answer>1</answer>

(example2) REAL

Image



Analysis view

HEURISTIC

**<think>A street scene featuring a woman walking along the roadside, vehicles in motion, traffic lights, and commercial signage including a 'Farmax' sign and a gas station billboard. The overall layout suggests an urban or suburban roadway with mixed pedestrian and vehicular activity under daylight conditions.</think>
<skill_select>Region Analysis</skill_select>**

REASONING

<think>The cropped region reveals fine-grained realism: the woman's clothing exhibits natural, non-repeating fabric folds and lighting consistent with ambient daylight; her posture and stride are anatomically plausible with motion blur matching walking dynamics. The road surface shows realistic asphalt texture, faded lane markings with irregular edges, and natural wear. The curb and adjacent grass display organic imperfections — uneven growth, minor debris, and discoloration — difficult to synthesize. Lighting and shadows are coherent across global and local views, with the woman's shadow correctly aligned to the scene's light source. No synthetic artifacts such as smeared edges, texture repetition, or perspective mismatches are present. All evidence supports the image being REAL.</think>
<answer>0</answer>

Figure 16: **Region Analysis.** Skill-based forensic analysis focusing on localized scene regions and object-level details to identify potential inconsistencies in local structures, textures, and semantic compositions.

Skill 9: Animal Anatomy

(example1) FAKE

Image

HEURISTIC

<think>A small rodent resembling a mouse or vole with prominent bat-like wings attached to its back, positioned on grass and nibbling a blade of grass with front paws. Wings are fully extended, membranous, with visible skeletal support similar to a bat. </think><skill_select>Animal Anatomy</skill_select>

REASONING

<think>The image is synthetic or digitally manipulated because rodents do not possess wings; the anatomical fusion of bat wings with rodent physiology violates known biological constraints, including skeletal structure, musculature, and evolutionary adaptation. While the rodent's body and behavior appear realistic, the wings render the entity biologically impossible.</think><answer>1</answer>

(example2) REAL

Image

HEURISTIC

<think>The main feature is a Highland cow standing in a grassy field, facing the camera. It has long, shaggy reddish-brown fur, prominent curved horns, and visible ear tags. The background includes blurred trees and distant hills under an overcast sky.</think><skill_select>Animal Anatomy</skill_select>

REASONING

<think>The depiction is realistic: anatomy, proportions, horn structure, fur texture, and ear tags align with known biological norms for Highland cattle. Posture and behavior are plausible for a domesticated bovine in a pasture. No synthetic or impossible features are present.</think><answer>0</answer>

Figure 17: **Frequency Consistency.** Skill-based forensic analysis that incorporates frequency-spectrum analysis models to examine whether the image exhibits realistic frequency distributions and spectral consistency patterns.

Skill 10: Frequency Consistency

(example1) FAKE

Image

HEURISTIC

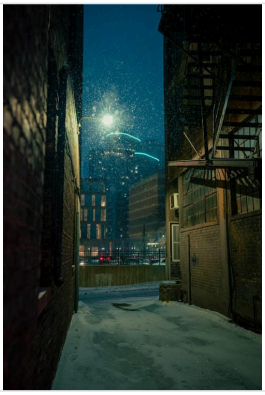
<think>A close-up of a cat's face with white and orange fur. The cat has greenish-yellow eyes, visible whiskers, and pointed ears. The fur around the ears and forehead shows fine strands. The background is dark and out of focus. Lighting appears to come from the front-right, casting highlights on the right side of the face and creating shadows on the left. The nose is pink, and the mouth area is subtly defined. The fur texture varies across regions, with some areas appearing smoother than others. </think><skill_select>Frequency Consistency</skill_select>

REASONING

<tool>1</tool><think>The image exhibits unnatural frequency characteristics. The fur texture, particularly around the cheeks and forehead, lacks the expected high-frequency noise and randomness typical of real animal fur. Instead, it displays overly smooth, airbrushed regions with uniform texture that do not vary naturally with lighting or depth. The edges of the whiskers and ear tips are unnaturally sharp and consistent in thickness, lacking the subtle tapering and micro-irregularities expected in real photographs. Additionally, the transition between lit and shadowed fur regions shows no natural high-frequency detail loss due to occlusion or depth, suggesting frequency-domain manipulation. These anomalies are inconsistent with real-world imaging processes and indicate synthetic generation.</think><answer>1</answer>

(example2) REAL

Image



HEURISTIC

<think>A narrow alleyway at night with snow on the ground. Brick walls on both sides, one with a fire escape and windows. Snowflakes are visible in the air, illuminated by a bright overhead light. In the background, a city skyline is visible with lit buildings, including one with a curved neon light. The ground shows footprints and tire tracks in the snow. Shadows are cast by the walls and structures. The sky is dark blue.</think><skill_select>Frequency Consistency</skill_select>

REASONING

<tool>0</tool><think>The image exhibits natural high-frequency noise in the snowflakes and brick textures, with irregular distribution and varying sizes that match real-world atmospheric and surface conditions. Edges of the fire escape, windows, and snow-covered ground show consistent sharpness appropriate for a long-exposure night shot, without artificial sharpening halos. The texture of the brick walls shows micro-variations and weathering, not repetitive patterns. The snowfall contains fine, non-uniform particles with natural motion blur and depth variation, consistent with real camera capture. Lighting gradients and shadow transitions contain subtle frequency modulations that align with real optical systems. No over-smoothed regions or unnatural frequency suppression are present. </think><answer>0</answer>

Figure 18: **Frequency Consistency.** Skill-based forensic analysis that incorporates frequency-spectrum analysis models to examine whether the image exhibits realistic frequency distributions and spectral consistency patterns.

Skill 11: Pixel Consistency

(example1) FAKE

Image

HEURISTIC

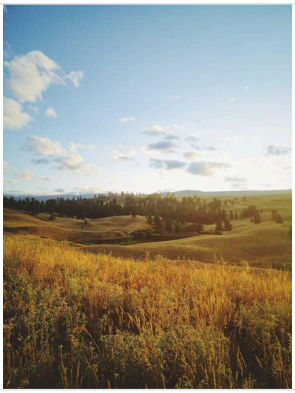
<think>A person with short brown hair is facing forward, wearing a light blue denim jacket over a white t-shirt, paired with blue jeans and a black belt. The background consists of vertical strings of out-of-focus circular lights in warm yellow and cool blue tones, creating a bokeh effect. The person's skin appears smooth, and the denim texture is visible but uniform. Lighting on the face is even, with soft shadows under the chin and around the nose. No visible camera artifacts, dust, or sensor noise are present in any region.</think><skill_select>Pixel Consistency</skill_select>

REASONING

<tool>1</tool><think>The image exhibits abnormal pixel-level noise characteristics. The skin texture is unnaturally smooth with no visible pore structure or micro-texture, inconsistent with real camera capture under soft lighting. The denim fabric shows uniform, repetitive stitching and fading patterns without the expected local variations in thread density or wear. The background bokeh lights display perfectly circular, evenly distributed blur with no chromatic aberration or sensor noise variation across the frame. The transition between the subject and background lacks natural edge noise or depth-of-field gradient artifacts. These pixel-level inconsistencies — including over-smoothed regions, lack of sensor noise, and uniform texture rendering — are inconsistent with real camera imaging pipelines and indicate synthetic generation.</think><answer>1</answer>

(example2) REAL

Image



HEURISTIC

<think>The image shows a wide landscape under a bright sky with scattered clouds. In the foreground, there is tall golden grass with visible individual blades and some green shrubs. The midground consists of rolling hills covered in green and golden vegetation, with clusters of dark trees. The background includes distant tree lines and faint mountain silhouettes under a clear blue sky. Lighting appears to come from the upper right, casting soft shadows and warm highlights across the terrain. No human-made objects or text are visible.</think><skill_select>Pixel Consistency</skill_select>

REASONING

<tool>0</tool><think>Pixel-level analysis reveals natural, non-uniform texture noise across the grass in the foreground, consistent with real camera sensor noise. Individual blades of grass show fine, irregular edges and subtle color variation without over-smoothing. The transition between lit and shadowed areas in the hills exhibits gradual tonal shifts and micro-contrast consistent with optical capture. Clouds show granular, non-repeating noise patterns typical of atmospheric scattering captured by a camera. No uniform denoising artifacts, unnatural sharpening halos, or repetitive pixel patterns are present. Local pixel statistics, including luminance variance and chroma noise distribution, align with those expected from a real photographic capture under natural lighting conditions.</think><answer>0</answer>

Figure 19: **Pixel Consistency.** Skill-based forensic analysis that leverages pixel-level forensic models to examine local pixel distributions, texture continuity, and low-level visual inconsistencies.

Skill 12: Transformation Consistency

(example1) FAKE

Image

Analysis view

HEURISTIC

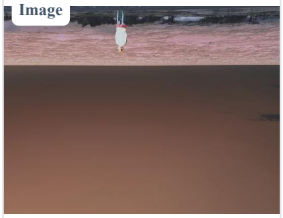
<think> The image appears to have undergone a geometric rotation transformation. The human subject and surrounding buildings are oriented sideways relative to the natural viewing direction, indicating that the current image orientation is inconsistent with normal real-world gravity alignment. The scene requires rotational correction to restore a physically plausible orientation, where the person stands upright and the architectural structures align vertically."transform_params": {"rotate": 90, "flip_h": true, "flip_v": true, "invert": false}</think><skill_select>Transformation Consistency </skill_select>

REASONING

<think>After restoration and consistency checking, the image shows signs of being synthetic. The face appears overly smooth and oily, lacking real skin texture and pores, with unnatural lighting consistency and softened details that deviate from physically plausible real-world imaging.</think><answer>1</answer>

(example2) REAL

Image



Analysis view

HEURISTIC

<think>The image displays a scene with objects appearing vertically mirrored — for example, text or structural elements that would normally be upright now appear upside down. Additionally, the color palette is inverted: bright areas appear dark and dark areas appear bright, suggesting a negative color effect. Spatial relationships between objects remain intact, but their orientation and tonal values are altered. No rotation is evident — the horizon or structural lines remain aligned with the image frame.", "transform_params": {"rotate": 0, "flip_h": false, "flip_v": true, "invert": true}</think><skill_select>Transformation Consistency</skill_select>

REASONING

<think>After restoration and consistency analysis, the image appears fully consistent with real-world photography. Both the human subject and the ocean scene exhibit natural texture, realistic lighting, and physically plausible details, indicating strong alignment with genuine image formation.</think><answer>0</answer>

Figure 20: **Transformation Consistency.** Skill-based forensic analysis that evaluates robustness under geometric rotation and color transformations, examining whether structural layout and chromatic relationships remain consistent and physically plausible after such changes.