REVEAL: Reasoning-Enhanced Forensic Evidence Analysis for Explainable AI-Generated Image Detection

Huangsen Cao[1,⋆], Qin Mei[1], Zhiheng Li[1], Yuxi Li[2], Zhan Meng[1], Ying Zhang[2], Chen Li[2], Zhimeng Zhang[1], Xin Ding[3], Yongwei Wang[1,✉], Jing LYU[2], and Fei Wu[1]

[1] Zhejiang University, China
[2] WeChat Vision, Tencent Inc
[3] Nanjing University of Information Science and Technology, China
yongwei.wang@zju.edu.cn

Abstract. The rapid progress of visual generative models has made AI-generated images increasingly difficult to distinguish from authentic ones, posing growing risks to social trust and information integrity. This motivates detectors that are not only accurate but also forensically explainable. While recent multimodal approaches improve interpretability, many rely on post-hoc rationalizations or coarse visual cues, without constructing verifiable chains of evidence, thus often leading to poor generalization. We introduce REVEAL-Bench, a reasoning-enhanced multimodal benchmark for AI-generated image forensics, structured around explicit chains of forensic evidence derived from lightweight expert models and consolidated into step-by-step chain-of-evidence traces. Based on this benchmark, we propose REVEAL (Reasoning-enhanced Forensic Evidence Analysis), an explainable forensic framework trained with expert-grounded reinforcement learning. Our reward design jointly promotes detection accuracy, evidence-grounded reasoning stability, and explanation faithfulness. Extensive experiments demonstrate significantly improved cross-domain generalization and more faithful explanations to baseline detectors. All data and source codes are publicly available at: https://github.com/TrustMedia-zju/REVEAL.

Keywords: AI-Generated Image Detection · Reasoning-Enhanced Forensic Benchmark · Explainable Image Forensics

1 Introduction

With the rapid evolution of generative artificial intelligence techniques, synthesized images have reached a level of visual realism that can readily deceive human perception [6,11,21]. While these technologies unlock substantial creative

✉ Corresponding author.
⋆ Work done during an internship at WeChat Vision, Tencent Inc.

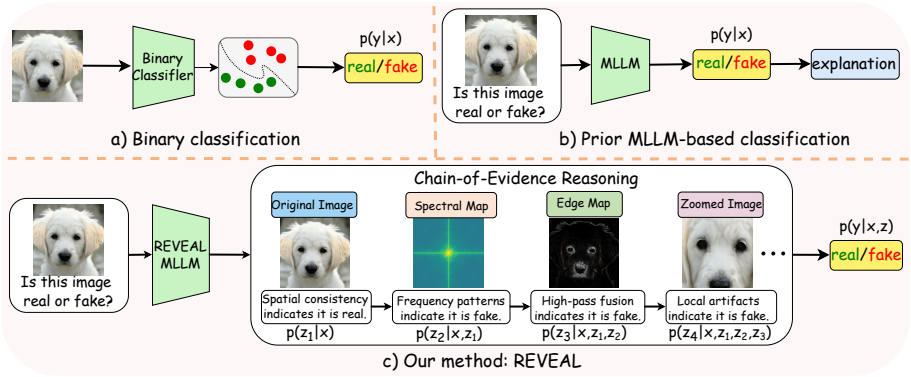

Fig. 1: (a) Conventional binary classification-based detection. (b) Prior MLLM-based post-hoc rationalization approaches, where explanations are generated *after* the detection decision. (c) Our REVEAL framework, enabling reasoning-enhanced forensic analysis through multi-view forensics and explicit chain-of-evidence reasoning.

and economic value in digital art, design, and film production, they also raise serious concerns regarding misinformation, privacy violations, and copyright issues. Continual advances in advanced diffusion models such as FLUX [23] and SDv3.5 [8], further exacerbate the difficulty of distinguishing real from synthetic content, making reliable detection an urgent research priority.

Recent research [4,33,37,47,52,53] has made significant progress in detecting AI-generated images. However, most existing detection methods are primarily optimized for binary classification and provide limited support for forensic analysis. Conventional detectors based on discriminative classifiers often provide no explanation beyond a label (e.g. [52, 57]). More interpretable designs (e.g. rule-based models or decision trees) can expose partial decision interpretation, but they frequently generalize poorly (see Table 6) and remain strongly coupled to the underlying detector's feature biases.

Multimodal large language models (MLLMs) offer a promising direction by combining visual perception with language-based reasoning [19, 59]. Recent works, including GPT-4-based detection [17], AIGI-Holmes [63], FakeBench [31], and RAIDX [29], attempt to improve explainability by generating human-readable rationales. However, as illustrated in Figure 1, current MLLM-based approaches exhibit two key limitations. First, explanations are generally *post hoc*, i.e., explanations are produced *after* prediction without an explicit mechanism ensuring that intermediate reasoning steps are supported by verifiable evidence. Second, MLLMs are typically used as *general-purpose* visual classifiers that detect high-level anomaly patterns (e.g. unnatural lighting or blurred boundaries), rather than as evidence-centric forensic systems that systematically collect, analyze, and synthesize specialized evidence into an auditable decision process.

A major cause of these limitations is the lack of datasets and training objectives that support forensic explainability and enforce evidence-grounded reason-

ing. Existing benchmarks often provide only image-level labels or brief textual justifications (e.g. FakeBench [31]), which do not capture the structured evidence required for forensic analysis. Similarly, explanation methods based on vanilla MLLMs or retrieval-augmented prompting (e.g. RAIDX [29]) may generate fluent explanations, but these remain weakly grounded when not explicitly tied to image-specific chains of forensic evidence.

These limitations point to two key challenges for reasoning-enhanced synthetic image detection: (1) the lack of reasoning-oriented forensic datasets, where annotations include structured evidence and step-by-step support for a final judgment; and (2) limited reasoning-based explainability, since current MLLM-based detectors often generate rationales that are not explicitly verifiable, often leading to limited generalization and unreliable forensic claims.

To address these challenges, we introduce REVEAL-Bench, a reasoning-oriented *benchmark* for AI-generated image forensics. Unlike existing approaches that emphasize learning generic visual correlations, our pipeline is designed around *expert-grounded evidence analysis*. For each image, we employ eight lightweight expert models to extract structured low-level forensic evidence. This evidence is then provided to a large model to produce a chain-of-evidence (CoE) annotation that links concrete cues to intermediate inferences and the final conclusion. By consolidating multi-round analyses from specialized experts into a single structured CoE trace, REVEAL-Bench enables verifiable forensic reasoning where high-level decisions are explicitly supported by low-level evidence.

Building on REVEAL-Bench, we propose the REVEAL *framework*, a two-stage training paradigm that enforces evidence-grounded forensic reasoning in MLLMs. In the first stage, we employ supervised fine-tuning to teach the MLLM a canonical CoE structure. In the second stage, we introduce R-GRPO (Reasoning-enhanced Group Relative Preference Optimization), an expert-grounded policy optimization method with a novel reward design that promotes reliable forensic reasoning and verifiability of forensic analysis. Concretely, R-GRPO jointly optimizes detection accuracy, reasoning stability, and multi-view consistency, encouraging the model to synthesize explicit evidence into a coherent decision trace rather than relying on superficial pattern matching. This yields the REVEAL detector that is more generalizable and faithful in forensic analysis.

In summary, our key contributions are threefold:

- We introduce REVEAL-Bench, a reasoning-enhanced dataset for explainable AI-generated image detection. Unlike prior detection datasets that rely on post-hoc explanations, REVEAL-Bench is structured around expert-grounded and verifiable forensic evidence, integrating an explicit chain-of-evidence under an evidence-then-reasoning paradigm.
- We propose the REVEAL framework, a novel progressive two-stage training paradigm that instills standardized and evidence-grounded reasoning in MLLMs. Its core component, R-GRPO, optimizes the evidence synthesis capability of MLLMs to jointly improve detection accuracy, cross-domain generalization, and explanation faithfulness.

– Our approach achieves stronger detection performance, improved generalization, and higher explanation fidelity, establishing a new state of the art for evidence-grounded reasoning in image forensics.

2 Related Work

Conventional AI-Generated Image Detection. The rapid evolution of generative models, e.g., GANs [9, 11], autoregressive models [49], diffusion-based models [8,12,14,16,39,42], has driven AI-generated images to near-photorealistic quality, challenging conventional detection methods. Early forensic studies focused on traditional manipulations like splicing or copy-move, analyzing noise inconsistencies, boundary anomalies, or compression artifacts [62]. Researchers then shifted focus to generation artifacts, such as up-sampling grid effects, texture mismatches, or abnormal high-frequency decay [7, 10, 35]. For example, the Spectral Learning Detector [20] models the spectral distribution of authentic images, treating AI-generated samples as out-of-distribution anomalies, achieving consistent detection across generators. However, as generators incorporate post-processing techniques like super-resolution, these low-level statistical clues become increasingly subtle and less reliable for robust detection.

Recent methods employ general-purpose feature extractors, such as CNN- or ViT-based detectors, to learn discriminative features directly. While lightweight CNNs achieve strong benchmark performance [25], methods like the Variational Information Bottleneck (VIB) network [60] aim to enhance generalization by constraining feature representations through the information bottleneck principle to retain only task-relevant information. Post-hoc Distribution Alignment (PDA) [50] attempts to improve robustness to unseen generators by aligning regenerated and real distributions to detect unseen generators. Recently, NPR [47] has become a representative approach by capturing low-level artifacts, demonstrating strong generalization capability. Similarly, HyperDet [3] and AIDE [57] achieve robust generalization through high-frequency spectrum analysis. Despite their discriminatory power, these approaches remain limited in forensic value, as their conclusions rely on global statistics and lack the semantic, verifiable evidence required for comprehensive explainability.

MLLM-based AI-generated Image Detection. The emergence of MLLMs [32, 51] has accelerated the development of explainable image forensics by leveraging their advanced cross-modal understanding [45, 55]. Early efforts reformulated detection as a visual question answering task [5, 17, 22], allowing MLLMs to provide accompanying descriptive text. FatFormer [33] extended it with a forgery-aware adapter to improve generalization on CLIP-ViT [38] encoder.

Subsequent studies focused on constructing task-specific multimodal datasets for fine-tuning. FakeBench [31] and LOKI [58] provide synthetic images with manually written, high-level forgery descriptions. Holmes-Set [63] utilized small models for initial image filtering and a multi-expert jury mechanism to generate post-hoc explanatory texts. At the methodological level, FakeShield [56], ForgerySleuth [43], ForgeryGPT [34] and SIDA [15] fine-tune MLLMs to achieve

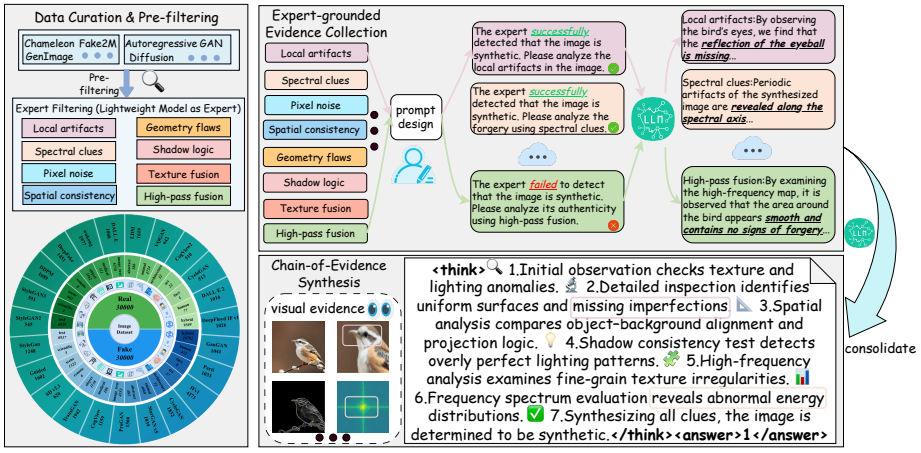

Fig. 2: The pipeline of REVEAL-Bench construction. This figure illustrates our data processing pipeline, which consists of three stages: Data Curation & Pre-filtering, Expert-grounded Evidence Collection, and Chain-of-Evidence (CoE) Synthesis.

explainable forgery detection and localization. AIGI-Holmes [63] integrates low-level visual experts with reasoning modules. RAIDX [29] combines retrieval-augmented generation (RAG) [26] with GRPO to improve text description. More recently, Ivy-Fake [18] presents a unified explainable framework and benchmark for both image and video AIGC detection, while Veritas [48] explores pattern-aware reasoning to enhance the generalization capability of deepfake detection.

However, existing datasets and methods still suffer from two key limitations: First, the explanations are attributed to post-hoc rationalizations, often relying on the MLLM's general knowledge and visual classification capabilities, failing to achieve logical synthesis of specialized forensic evidence. Second, they often lack structured, fine-grained forensic evidence required to support a verifiable causal link between low-level artifacts and the final forensic judgments.

3 REVEAL-Bench

As illustrated in Figure 2, this study constructs the REVEAL-Bench dataset through a rigorous, three-stage pipeline for reasoning-based image forensic: Data Curation & Pre-filtering, Expert-grounded Evidence Collection, and Chain-of-Evidence (CoE) Synthesis. This approach is fundamentally distinct as it replaces manual and subjective labeling with a process that systematically integrates verifiable evidence from specialized models with the logical synthesis capabilities of large vision-language models. The resulting dataset contains explicit, expert knowledge-grounded Chain-of-Evidence annotations, which are crucial for training forensic detectors with superior transparency and generalization.

Data Curation & Prefiltering.

To ensure sufficient content, generator, and artifact diversity, we aggregate several prominent AI-generated detection benchmarks, including CNNDetection [52], UnivFD [37], AIGCDetectBenchmark [61], GenImage [64], Fake2M [36], and Chameleon [57]. This yielded an initial corpus of approximately 5,120K synthetic images and 850K authentic images. To manage annotation costs while ensuring high data quality, we implemented a stratified sampling strategy based on automated quality assessments [44] and image resolution. Specifically, we sampled images based on aesthetic scores (50% high, 30% medium, 20% low), and image resolution, high-resolution ($\geq$512×512) images at 50%, medium-resolution (384×384–512×512) images at 30%, and low-resolution (<384×384) images at 20%. Images were also semantically classified into 13 major categories (e.g. humans, architecture, artworks). After rigorous multi-stage filtering and preprocessing to eliminate non-representative or low-quality samples, we obtained a balanced corpus spanning diverse categories, resolutions, and visual qualities, forming a reliable foundation for subsequent expert annotation and reasoning supervision.

Expert-Grounded Evidence Collection. To enable fine-grained and verifiable forensic analysis, we design and employ a set of eight lightweight and specialized expert models [3,27,28,40,46,47], each dedicated to screening and localizing a distinct category of synthetic artifact (detail in Appendix A). All experts are implemented using publicly available codebases and are directly adopted without additional training or fine-tuning, ensuring strong reproducibility and practical deployability. Specifically, these experts operate on complementary visual and signal representations. Local artifacts [28] focus on local artifact enhancement, while spectral clues [46] perform frequency spectrum analysis. Pixel noise [47] examines neighbor-pixel residuals, and spatial consistency [40] detects obvious fake cues. Geometry flaws [40] conducts object spatial analysis, whereas shadow logic [40] verifies shadow coherence. Finally, texture fusion [27] applies texture-frequency fusion, and high-pass fusion [3] performs high-pass semantic fusion. In addition, object segmentation cues are used to verify semantic boundary coherence, while texture-focused and high-pass frequency representations further expose unnatural smoothness or repetitive patterns commonly introduced by generative models. This is a crucial distinction from prior work, such as AIGI-Holmes [63], which uses experts primarily for global filtering. Our experts, by contrast, provide structured, machine-readable evidence, including artifact masks and diagnostic labels. These eight outputs constitute the necessary forensic evidence foundation. By conditioning the LVLM on these high-fidelity, structured references, we ensure the final generated explanations are faithful, logically consistent, and verifiable against objective, low-level artifact data. This expert-grounded decompositional analysis effectively bridges the gap between small-model perception of artifacts and large-model logical reasoning.

Chain-of-Evidence Synthesis. As shown in Figure 2, after the specialized expert annotation, the initial eight rounds of multi-perspective diagnostic outputs are diverse and fragmented. To construct a unified and progressive reasoning dataset suitable for Chain-of-Thought (CoT) fine-tuning, we leverage a

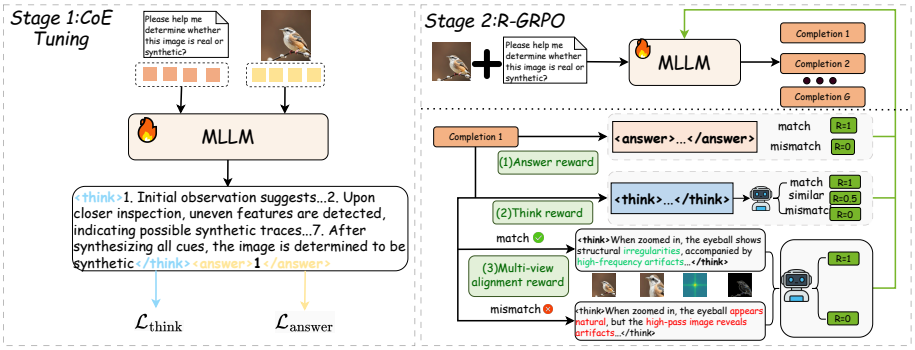

Fig. 3: Overview of the proposed REVEAL framework. The pipeline mainly consists of two stages: CoE Tuning and R-GRPO.

high-capacity LVLM (Qwen-2.5VL-72B [1]) to perform structured knowledge consolidation. This process reconstructs the diverse, specialized evidence into a single, cohesive, and auditable reasoning trace, formatted using a standard $<think>\cdots</think>\cdot<answer>\cdots</answer>$ structure. Starting from a large-scale pool of approximately 6M annotated samples, we conduct stringent quality filtering and reasoning trace refinement, ultimately distilling a curated 60K high-quality explainable dataset (30K real, 30K synthetic) with structured Chain-of-Evidence (CoE) supervision—corresponding to a selection rate of roughly 1%.

Fundamentally different from existing datasets like AIGI-Holmes [63] and FakeBench [31], which merely provide generic explanations, REVEAL-Bench explicitly formalizes the link between low-level expert evidence and high-level judgments. This two-stage pipeline transforms the detection tasks into a reasoning task, offering coherent CoE annotations that enhance logical consistency, minimizing annotation noise, and supporting supervision paradigms with reinforcement learning techniques to improve explanation fidelity and generalization.

4 Methodology

4.1 Overview of REVEAL Framework

As illustrated in Figure 3, the overall training pipeline adopts a two-stage progressive training paradigm inspired by advanced policy optimization-based reinforcement learning techniques [13]. It is worth mentioning that the eight expert models are used only for offline dataset construction to generate artifact annotations and diagnostic labels, and do not participate in model training or inference.

We first perform supervised fine-tuning (SFT) on a consolidated Chain-of-Evidence (CoE) dataset to obtain a base policy that can deduce the required forensic reasoning procedure. While this stage establishes the fundamental reasoning-based forensic structure, the resulting model still exhibits limitations

in logical consistency, forensic accuracy, and robustness. To mitigate these limitations, we propose a novel reinforcement learning algorithm: *Reasoning-enhanced Forensic Evidence Analysis (R-GRPO)*. R-GRPO extends standard Group Relative Policy Optimization (GRPO) by incorporating a task-specific composite reward that dynamically aligns forensic reasoning trajectories and stabilizes policy updates, significantly enhancing semantic consistency and reasoning robustness.

4.2 Reasoning-Enhanced Progressive Multimodal Training

We introduce REVEAL (Reasoning-enhanced Forensic Evidence AnaLysis), a progressive multimodal training framework comprising two sequential stages designed to facilitate logically consistent and verifiable forensic reasoning in multimodal models.

Stage 1: Chain-of-Evidence Tuning (CoE Tuning). In the initial stage, we perform cold-start supervised fine-tuning to establish a stable, stepwise reasoning policy and a consistent output paradigm built upon the REVEAL-Bench dataset. Let x denote the visual input, $z = (z_1, \ldots, z_T)$ denote the tokenized reasoning sequence (Chain-of-Evidence, CoE), and y denote the final classification label. We adopt an explicit joint reasoning–decision modeling paradigm, where the final prediction y is conditioned on the explicit reasoning trace z. This formulation enforces a *think-then-answer* mechanism, fundamentally distinct from post-hoc rationalizations (i.e. modeling $p(y \mid x)$ and then $p(z \mid x, y)$), thereby achieving causally grounded genuine explanations.

Concretely, we factorize the joint conditional probability as

$$p(y, z \mid x) \;=\; p(z \mid x)\, p(y \mid x, z), \tag{1}$$

which structurally encourages the model to first generate verifiable reasoning evidence and subsequently derive the final prediction conditioned directly on that reasoning process.

Maximizing the likelihood under (1) corresponds to minimizing the following negative log-likelihood loss:

$$\mathcal{L}_{\mathrm{NLL}}(x, y, z; \theta) \;=\; -\log p_\theta(z \mid x) \;-\; \log p_\theta(y \mid x, z). \tag{2}$$

For training control and to explicitly balance the emphasis on reasoning quality versus final decision accuracy, we decompose $\mathcal{L}_{\mathrm{NLL}}$ into two components, the reasoning generation loss $\mathcal{L}_{\mathrm{think}}$ and the answer loss $\mathcal{L}_{\mathrm{answer}}$,

$$\mathcal{L}_{\mathrm{think}} \;=\; -\sum_{t=1}^{T} \log p_\theta(z_t \mid z_{<t}, x), \tag{3}$$

$$\mathcal{L}_{\mathrm{answer}} \;=\; -\log p_\theta(y \mid x, z), \tag{4}$$

We then employ a weighted composite SFT loss:

$$\mathcal{L}_{\mathrm{SFT}} = (1-\alpha)\,\mathcal{L}_{\mathrm{think}} + \alpha\,\mathcal{L}_{\mathrm{answer}} + \eta\,\mathrm{KL}\big(\pi_{\mathrm{pre}} \| \pi_\theta\big). \tag{5}$$

Here, $\alpha \in (0, 1)$ balances the contribution of the reasoning trace and the answer loss. In this work, α is set to 0.1 to ensure that reasoning serves as an auxiliary signal while prioritizing the correctness of the final answer. The KL regularization term keeps the fine-tuned policy π_θ close to the pretrained policy π_{pre}, effectively mitigating catastrophic forgetting.

Stage 2: Reasoning-enhanced Group Relative Policy Optimization (R-GRPO).

Group Relative Policy Optimization (GRPO) [41] improves training stability by replacing the traditional neural-network critic with a group-based baseline. For a given input x, we sample a group of G trajectories $\{\tau_i\}_{i=1}^{G}$ from the current policy π_θ. Instead of relying on a separate value model, GRPO computes the advantage A_i by standardizing the rewards within the group, effectively rewarding outputs that outperform the group average:

$$A_i = \frac{r_i - \text{mean}(r_1, r_2, \ldots, r_G)}{\text{std}(r_1, r_2, \ldots, r_G) + \epsilon}, \tag{6}$$

where r_i is the reward for trajectory τ_i, and ϵ is a small constant to avoid division by zero.

The optimization objective maximizes the group-relative advantage using a PPO-style surrogate loss while regularizing the update with a KL divergence penalty against a reference policy π_{ref} (typically the initial SFT model) to prevent catastrophic forgetting:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_x \left[\frac{1}{G} \sum_{i=1}^{G} \Big(\min \Big(\frac{\pi_\theta(\tau_i|x)}{\pi_{\text{old}}(\tau_i|x)} A_i, \text{clip}\Big(\frac{\pi_\theta(\tau_i|x)}{\pi_{\text{old}}(\tau_i|x)}, 1-\epsilon, 1+\epsilon \Big) A_i \Big) \Big) \right] - \beta D_{KL}(\pi_\theta \| \pi_{\text{ref}}) \tag{7}$$

where β controls the strength of KL regularization, π_{old} is the policy before the current update, and $\text{clip}(\cdot, 1-\epsilon, 1+\epsilon)$ ensures stable updates by limiting the probability ratio.

Reasoning-Enhanced GRPO (R-GRPO). To employ GRPO for forensic analysis tasks, we propose *R-GRPO*, which augments the objective with a task-aware composite reward specifically designed to capture forensic fidelity and reasoning robustness. Similar to recent GRPO variants [30, 48], which improve optimization through reward design or policy refinement, R-GRPO further introduces forensic-oriented reward modeling for evidence-based reasoning. Let y denote the generated answer, y^* the reference answer, $z = (z_1, \ldots, z_T)$ the reasoning tokens, and $\{v_m(x)\}_{m=1}^{M}$ a set of multi-view visual evidence corresponding to the visual analysis perspectives used by the expert models during dataset construction (e.g., spectral representations, high-pass filtered images, edge responses, and localized artifact regions).

Rationale for Agent-based Reward Modeling. In preliminary experiments, we observed that simple metric-based rewards (e.g. computing r_{sem} via cosine similarity of sentence embeddings) fail to adequately capture the semantic coherence and contextual logic required for high-quality forensic explanations. To

address this limitation, we employ a pretrained large vision-language model, Qwen-3-VL-8B, as an intelligent agent (*Agent*) to evaluate and score model responses, without requiring any additional training or task-specific design. This Agent-based assessment considers contextual logic, explanation coherence, and factual consistency against the provided structured evidence, thereby generating a more human-aligned and explainable reward signal than purely metric-based approaches, with additional validation of assessor reliability and bias analysis provided in Appendix F.

R-GRPO defines three complementary, evidence-driven reward components:

(1) Answer Reward r_{ans}. This binary reward captures only the correctness of the response with respect to the ground-truth answer.

$$r_{\text{ans}}(y, y^*) = \begin{cases} 1, & \text{if } y = y^*, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

(2) Think Reward r_{think}. This reward quantifies the quality and structural integrity of the reasoning trace z.

Let $z = (z_1, \ldots, z_T)$ be the generated reasoning trace and $z^* = (z_1^*, \ldots, z_{T^*}^*)$ the ground-truth reasoning trace (when available). Define a perturbed trace $\tilde{z} = \text{shuffle}(z)$. Then

$$r_{\text{think}}(z, z^*, \tilde{z}) \;=\; \mathcal{A}_{\text{sem}}(z, z^*) + \mathcal{A}_{\text{logic}}(z, \tilde{z}), \tag{9}$$

where $\mathcal{A}_{\text{sem}}$ measures alignment between the generated and reference reasoning, and $\mathcal{A}_{\text{logic}}(z, \tilde{z})$ evaluates the logical coherence of the trace. Here, both $\mathcal{A}_{\text{sem}}$ and $\mathcal{A}_{\text{logic}}$ are computed by a pretrained large language model agent, and are applied solely to evaluate the content within the *<think>*$\cdots$*</think>* block. Crucially, $\mathcal{A}_{\text{logic}}$ evaluates logical coherence by penalizing the model if minor structural perturbations $\tilde{z}$ severely alter the inferred conclusion. This mechanism encourages the model to maintain sequential consistency and ensures that the reasoning steps are robustly connected.

(3) Multi-view Alignment Reward r_{view}. This reward encourages the generated reasoning trace z to be robustly grounded in evidence that persists across different forensic views of the image.

$$r_{\text{view}}(z, x) \;=\; \mathcal{A}_{\text{view}}\Big(z, \{v_m(x)\}_{m=1}^{M}\Big), \tag{10}$$

where $\mathcal{A}_{\text{view}}$ is computed by a pretrained agent and evaluates how well the content within the *<think>*$\cdots$*</think>* block aligns with visual evidence under various transformations (e.g., spectral, high-pass). This ensures that each view is analyzed correctly and accurately while promoting cross-artifact generalization.

The composite trajectory reward $R(\tau)$ combines these terms:

$$\begin{aligned} R(\tau) = \lambda_a r_{\text{ans}}(y, y^*) + \lambda_t r_{\text{think}}(z, z^*, \tilde{z}) \\ + \lambda_v r_{\text{view}}(z, x), \end{aligned} \tag{11}$$

where $\lambda_a, \lambda_t, \lambda_v \geq 0$ are tunable parameters balancing the contributions of the three rewards. In our experiments, we set $(\lambda_a, \lambda_t, \lambda_v) = (0.8, 0.1, 0.1)$, ensuring that the reasoning rewards consistently serve to support accurate final predictions. For improved stability, rewards are standardized within each sampled group and directly used as the group-relative advantage:

$$\widehat{A}_i = \frac{R(\tau_i) - \mu_{\text{group}}}{\sigma_{\text{group}}}, \tag{12}$$

where μ_{group} and σ_{group} denote the mean and standard deviation of rewards within the group, respectively.

Unified GRPO with the R-GRPO objective. Combining the original GRPO formulation (7) with the R-GRPO composite reward (12), the unified optimization objective becomes

$$\mathcal{J}_{\text{R-GRPO}}(\theta) = \mathbb{E}_x \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min \Big(\frac{\pi_\theta(\tau_i|x)}{\pi_{\text{old}}(\tau_i|x)} \widehat{A}_i, \text{clip}\Big(\frac{\pi_\theta(\tau_i|x)}{\pi_{\text{old}}(\tau_i|x)}, 1-\epsilon, 1+\epsilon \Big) \widehat{A}_i \Big) \Bigg] - \beta D_{KL}(\pi_\theta \| \pi_{\text{ref}}) \tag{13}$$

where $\widehat{A}_i$ encodes both the group-relative comparison and the reasoning-enhanced composite reward.

These evidence-enhanced reward signals can effectively guide the model to optimize its reasoning trajectories, enforcing both stability and logical coherence in verifiable forensic evidence analysis.

5 Experiments

5.1 Experimental Settings

To comprehensively evaluate REVEAL, we conduct experiments on three datasets: REVEAL-Bench, GenImage [64], and REVEAL-Bench++ . REVEAL-Bench(see Table 1), a chain-of-evidence–annotated dataset for explainable synthetic image detection, serves as the in-domain dataset for training and evaluation. GenImage, a million-scale synthetic image dataset covering diverse representative generation methods, is used as an out-of-domain benchmark to assess generalization.

To further evaluate generalization to unseen generators, we additionally construct REVEAL-Bench++, a challenging test set of 10K images (2K per generator, 1K real and 1K AI-generated) produced by recent models, including FLUX [23], FLUX2 [24], Z-Image [2], Qwen-Image [54], and SDv3.5 [8], none of which appear in training. We train REVEAL on REVEAL-Bench and evaluate it on all three datasets (see Appendix B for training details).

Baselines. We compare REVEAL with state-of-the-art AI-generated image detection methods, including CNNSpot [52], UnivFD [37], NPR [47], HyperDet [3],

Table 1: Comparison with prior datasets. REVEAL-Bench is the first comprehensive reasoning dataset for synthetic image detection.

Dataset	#Image	Explanation	Multiview Fusion	Reasoning Process
CNNDetection [52]	720k	✗	✗	✗
GenImage [64]	1M	✗	✗	✗
FakeBench [31]	6K	✓	✗	✗
Holmes-Set [63]	69K	✓	✓	✗
REVEAL-Bench	60K	✓	✓	✓

Table 2: REVEAL improves over the strongest methods by **4.14%** on REVEAL-Bench and REVEAL-Bench++.

Method	REVEAL	SD3.5	FLUX	FLUX2	Qwen-Image	Z-Image	Mean
CNNSpot	87.80	71.70	70.30	61.70	82.35	59.10	72.16
UnivFD	86.95	84.45	84.55	83.65	85.85	66.75	82.03
NPR	95.40	53.00	51.20	51.60	51.40	53.60	59.37
HyperDet	93.25	88.80	79.70	65.20	88.15	70.80	80.98
AIGI-Holmes	93.10	82.14	79.35	76.41	75.43	69.47	79.32
Veritas	72.75	87.35	89.90	90.76	90.00	87.80	86.43
Ivy-xDetector	77.65	90.75	91.60	91.04	91.65	86.55	88.21
REVEAL	95.31	94.38	93.44	91.25	95.00	84.69	92.35

AIDE [57], VIB-Net [60], AIGI-Holmes [63], Veritas [48] and Ivy-xDetector [18]. For fair comparison, we retrain all baselines using their official codes under the same dataset splits and experimental protocol.

Evaluation Metrics. Following prior work, we report classification accuracy (ACC). ACC is the proportion of correctly classified samples over the full test set and measures overall detection correctness. Since REVEAL outputs textual predictions (Real/Fake), we map them to binary labels for computing ACC. Baseline methods follow the default decision thresholds in their official implementations. Moreover, since REVEAL produces texts rather than calibrated logits, we do not report metrics that require score outputs (e.g. average precision).

5.2 Generalizable Detection and Analysis

Table 2 reports performance comparisons on the in-domain dataset *REVEAL-Bench* and the challenging out-of-domain benchmark, *REVEAL-Bench++*, which contains images generated by FLUX, FLUX2, Z-Image, Qwen-Image, and SDv3.5. Table 3 reports results on another typical out-of-domain benchmark, *GenImage*. Overall, REVEAL achieves consistently strong generalization performances compared to baseline lightweight binary classifiers, maintaining high accuracy on both recent unseen generators (REVEAL-Bench++) and GenImage. We attribute this improvement to the CoE-based evidence collection and reasoning procedure, which reduces reliance on spurious domain-specific correlations.

On the in-domain setting *REVEAL-Bench*, smaller classifiers (e.g. NPR [47], AIDE [57]) can achieve highly competitive accuracy, likely due to their ability to fit dataset-specific statistical regularities. In contrast, REVEAL performs comparably on *REVEAL-Bench* while consistently outperforming these compact models under distribution shift, especially on the harder *REVEAL-Bench++* benchmark. These results suggest that while compact detectors remain attractive when computational efficiency and in-domain accuracy are primary concerns, reasoning-based forensic approaches like REVEAL offer substantially stronger robustness and generalization to unseen generators. Please refer to Appendices C–D for additional few-shot results and MLLM-based detector comparisons.

Table 3: Performance comparison of cross-domain generalization on GenImage dataset. REVEAL outperforms state-of-the-art baselines by at least **4.35%**.

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	*Mean*
CNNSpot [52]	62.45	74.25	73.85	63.55	73.60	73.70	71.35	39.45	66.53
UnivFD [37]	75.00	84.35	80.95	85.50	71.75	82.00	80.70	88.45	81.09
NPR [47]	84.80	88.85	88.05	85.10	**94.30**	87.05	84.45	88.95	87.69
HyperDet [3]	68.40	91.85	92.30	**100.0**	67.05	89.20	80.45	57.65	80.86
AIDE [57]	79.90	95.90	94.95	87.75	90.35	94.85	90.10	91.10	90.61
VIB-Net [60]	53.25	60.25	57.85	65.00	68.55	60.85	52.55	38.00	57.04
AIGI-Holmes [63]	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**

5.3 Ablation Studies

Impacts of Different MLLM Backbones. Our method is model-agnostic and can be applied to various multimodal large language models. We evaluate REVEAL using Qwen2.5-VL [1], LLaVA-1.5-VL [32], and Phi-3.5 as representative backbones on REVEAL-Bench. As shown in Table 4, REVEAL consistently achieves strong detection accuracy, indicating robust generalization across different architectures. We further observe that larger backbones generally yield better detection performance. This trend suggests that synthetic image detection within a reasoning-based framework benefits from model scaling, similar to other multimodal reasoning tasks. As the model capacity increases, the ability to synthesize forensic evidence improves accordingly.

Effectiveness of Reasoning-Oriented Training Strategies. We conduct ablations to analyze the impact of reasoning-oriented supervision and optimization. As shown in Table 5, we compare: (1) standard SFT without reasoning data (non-reasoning SFT); (2) CoE-based supervised fine-tuning (CoE Tuning); (3) an answer-first reasoning format, and (4) vanilla GRPO versus our proposed R-GRPO. Results show that incorporating reasoning data significantly improves detection performance. Models trained without reasoning supervision perform substantially worse, highlighting the importance of structured CoE annotations. Moreover, R-GRPO further enhances performance compared to vanilla GRPO, demonstrating the effectiveness of expert-grounded reward design in stabilizing and refining forensic reasoning.

5.4 Comparison with Existing Explainable Detectors

We compare REVEAL with two types of explainable detectors on REVEAL-Bench: conventional interpretable classifiers and AIGI-Holmes, a state-of-the-art MLLM-based detector. The former can be constructed from the predictions of eight expert models using rule-based aggregation methods such as majority voting or decision trees. As shown in Table 6, synthesizing expert predictions

Table 4: Detection performance across MLLM backbones with CoE Tuning and with Tuning plus R-GRPO.

Model	CoE Tuning	+ R-GRPO
Phi-3.5	83.75	87.19
Qwen2.5-VL-3b	87.18	89.06
Qwen2.5-VL-7b	85.73	92.19
llava-v1.5-7b	91.56	92.81
llava-v1.5-13b	93.06	95.31

Table 5: Ablation study of Answer-first Tuning, CoE Tuning, GRPO, and R-GRPO.

Answer-first	CoE	GRPO	R-GRPO	Acc
✗	✗	✗	✗	61.21
✓	✗	✗	✗	82.39
✗	✓	✗	✗	85.73
✗	✓	✓	✗	91.56
✗	✓	✗	✓	95.31

Table 6: Comparison with decision-based methods: majority voting and decision trees.

Method	Accuracy (%)
Best Lightweight Expert	65.48
Majority Voting	78.35
Decision Tree	74.75
REVEAL (Ours)	95.31

through a large-model-based CoE framework significantly improves detection accuracy compared with these rule-based approaches. While majority voting and decision trees capture coarse consensus signals, they lack the capacity to reason over nuanced forensic cues.

We further compare REVEAL with AIGI-Holmes in terms of explanation quality. Appendix E presents a human preference study in which REVEAL is preferred over prior methods by **26%**. Appendix G evaluates multi-view faithfulness, and Appendix H compares REVEAL with both closed- and open-source interpretability approaches, demonstrating consistent advantages.

5.5 Effectiveness of Expert-Guided CoE Annotations

To evaluate the effectiveness of expert guidance in dataset construction, we compare expert-guided annotations with those generated directly by an LLM without expert inputs. The comparison is conducted from three complementary perspectives (Figure 4). First, we evaluate annotation correctness. Expert-guided annotations exhibit substantially fewer labeling errors compared to direct LLM annotations, indicating reduced noise and improved reliability. Second, to assess the reliability of the explanations, we specifically evaluate GAN-generated images known to exhibit frequency-domain artifacts, examining whether the generated descriptions correctly identify abnormal signals (e.g. checkerboard artifacts). The expert-guided dataset consistently demonstrates more accurate and technically grounded explanations. Third, we conduct a human review of 100 samples involving specialized forensic terms. Expert-guided annotations achieve higher correctness and better terminology coverage than purely LLM-generated annotations. Overall, expert-guided CoE annotation significantly improves accuracy, interpretability, and domain alignment, which are critical for reliable training and evaluation in synthetic image forensics.

5.6 Robustness to Unseen Perturbations

We evaluate robustness against common post-processing perturbations on the REVEAL-Bench dataset. Specifically, we apply two typical post-processing operations to the original test images: Gaussian blur ($\sigma = 1, 2, 3, 4$) and JPEG

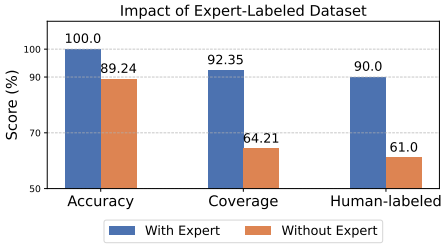

Fig. 4: Comparison of labeling strategies. Expert-guided annotations improve explanation accuracy and coverage.

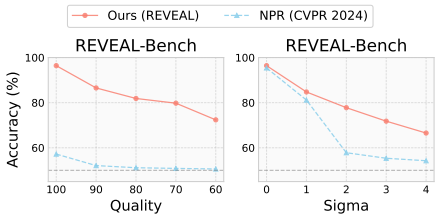

Fig. 5: Accuracy comparison between methods under various perturbation conditions.

compression (quality = 90, 80, 70, 60). For each distortion level, we compare REVEAL with the state-of-the-art baseline methods (see Figure 5). We can observe that REVEAL consistently maintains stronger performance under different perturbation severity, demonstrating improved robustness and better generalization across post-processing variations.

6 Conclusion

We presented REVEAL, a reasoning-enhanced framework for explainable AI-generated image detection. We introduced REVEAL-Bench, a dataset structured around expert-grounded forensic evidence with explicit chain-of-evidence annotations under an evidence-then-reasoning paradigm. Building on this benchmark, we proposed a two-stage training framework with R-GRPO, which guides multimodal LLMs to synthesize forensic evidence through structured reasoning, improving detection accuracy, generalization, and explanation fidelity. Extensive experiments demonstrate strong performance and robustness to unseen generators, advancing evidence-grounded reasoning for synthetic image forensics.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62502442, U24A20326, 62306147), Zhejiang Provincial Natural Science Foundation of China (No. LQN26F020009), and the Science-Tech Innovation Community of the Yangtze River Delta (No. 2023CSJZN0301).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923

2. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
3. Cao, H., Wang, Y., Liu, Y., Zheng, S., Lv, K., Zhang, Z., Zhang, B., Ding, X., Wu, F.: Hyperdet: Generalizable detection of synthesized images by generating and merging a mixture of hyper loras. arXiv preprint arXiv:2410.06044 (2024)
4. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: European conference on computer vision. pp. 103–120. Springer (2020)
5. Chang, Y.M., Yeh, C., Chiu, W.C., Yu, N.: Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. arXiv preprint arXiv:2310.17419 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Dzanic, T., Shah, K., Witherden, F.D.: Fourier spectrum discrepancies in deep network generated images. In: Advances in Neural Information Processing Systems (2020)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
10. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: Proceedings of the 37th International Conference on Machine Learning (2020)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
12. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
13. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239 (2020)
15. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., Cheng, G.: Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28831–28841 (2025)
16. Ji, Y., Yan, H., Lan, J., Zhu, H., Wang, W., Fan, Q., Zhang, L., Zhang, J.: Interpretable and reliable detection of ai-generated images via grounded reasoning in mllms. arXiv preprint arXiv:2506.07045 (2025)
17. Jia, S., Lyu, R., Zhao, K., Chen, Y., Yan, Z., Ju, Y., Hu, C., Li, X., Wu, B., Lyu, S.: Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4324–4333 (2024)

18. Jiang, C., Dong, W., Zhang, Z., Yu, F., Peng, W., Yuan, X., Bi, Y., Zhao, M., Zhou, Z., Si, C., et al.: Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. In: Proceedings of the 2026 International Conference on Multimedia Retrieval. pp. 2438–2447 (2026)
19. Jiang, T., Wu, L., Xia, S., Li, S., Yan, Z., Yang, H., Qiao, Y., Wang, Y.: Imagine before you predict: Interleaved latent visual reasoning for video event prediction. arXiv preprint arXiv:2606.05769 (2026)
20. Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E.: Any-resolution ai-generated image detection by spectral learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
22. Keita, M., Hamidouche, W., Bougueffa Eutamene, H., Taleb-Ahmed, A., Camacho, D., Hadid, A.: Bi-lora: A vision-language approach for synthetic image detection. Expert Systems **42**(2), e13829 (2025)
23. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (Jan 2024)
24. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
25. Lađević, L., Kramberger, T., Kramberger, R., Vlahek, D.: Detection of ai-generated synthetic images with a lightweight cnn. AI **5**(3), 76 (2024). `https://doi.org/10.3390/ai5030076`
26. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
27. Li, J., Jiang, W., Shen, L., Ren, Y.: Optimized frequency collaborative strategy drives ai image detection. IEEE Internet of Things Journal (2025)
28. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: An image transformation perspective. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. pp. 2405–2414 (2025)
29. Li, T., Huang, Z., Wen, H., He, Y., Lyu, S., Wu, B., Cheng, G.: Raidx: A retrieval-augmented generation and grpo reinforcement learning framework for explainable deepfake detection. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 11746–11755 (2025)
30. Li, X., Liang, Q., Li, Y., Zhang, Y., Li, C., Lyu, J., Lu, H., Jia, X.: Portrait-gen: Exemplar-driven grpo with dual-reward guidance for photorealistic portrait generation (2026), `https://arxiv.org/abs/2606.26930`
31. Li, Y., Liu, X., Wang, X., Lee, B.S., Wang, S., Rocha, A., Lin, W.: Fakebench: Probing explainable fake image detection via large multimodal models. IEEE Transactions on Information Forensics and Security (2025)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
33. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10770–10780 (2024)
34. Liu, J., Zhang, F., Zhu, J., Sun, E., Zhang, Q., Zha, Z.J.: Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. arXiv preprint arXiv:2410.10238 (2024)

35. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8057–8066 (2020)
36. Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., Ouyang, W.: Seeing is not always believing: Benchmarking human and model perception of ai-generated images. Advances in neural information processing systems **36**, 25435–25447 (2023)
37. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Lopes, R.G., Ayan, B.K., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494 (2022)
40. Sarkar, A., Mai, H., Mahapatra, A., Lazebnik, S., Forsyth, D.A., Bhattad, A.: 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. pp. 28140–28149 (2024)
41. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
43. Sun, Z., Jiang, H., Chen, H., Cao, Y., Qiu, X., Wu, Z., Jiang, Y.G.: Forgerysleuth: Empowering multimodal large language models for image manipulation detection. arXiv preprint arXiv:2411.19466 (2024)
44. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE transactions on image processing **27**(8), 3998–4011 (2018)
45. Talmor, A., Herzig, J., Lourie, N., Berant, J.: Commonsenseqa: A question answering challenge targeting commonsense knowledge. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4149–4158 (2019)
46. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5052–5060 (2024)
47. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: 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. pp. 28130–28139 (2024)
48. Tan, H., Lan, J., Tan, Z., Liu, A., Song, C., Shi, S., Zhu, H., Wang, W., Wan, J., Lei, Z.: Veritas: Generalizable deepfake detection via pattern-aware reasoning. arXiv preprint arXiv:2508.21048 (2025)
49. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems. vol. 30 (2017)

50. Wang, L., Chen, W., Li, Z., Guo, S.: Pda: Generalizable detection of ai-generated images via post-hoc distribution alignment. arXiv preprint arXiv:2502.10803 (2025)
51. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
52. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
53. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22445–22455 (2023)
54. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
55. Wu, T., Ma, K., Liang, J., Yang, Y., Zhang, L.: A comprehensive study of multimodal large language models for image quality assessment. In: European Conference on Computer Vision. pp. 143–160. Springer (2024)
56. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., Zhang, J.: Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761 (2024)
57. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for ai-generated image detection. arXiv preprint arXiv:2406.19435 (2024)
58. Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., et al.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models. arXiv preprint arXiv:2410.09732 (2024)
59. Zhang, F., Wu, L., Bai, H., Lin, G., Li, X., Yu, X., Wang, Y., Chen, B., Keung, J.: Humaneval-v: benchmarking high-level visual reasoning with complex diagrams in coding tasks. arXiv preprint arXiv:2410.12381 (2024)
60. Zhang, H., He, Q., Bi, X., Li, W., Liu, B., Xiao, B.: Towards universal ai-generated image detection by variational information bottleneck network. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23828–23837 (2025)
61. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: Patchcraft: Exploring texture patch for efficient ai-generated image detection. arXiv preprint arXiv:2311.12397 (2023)
62. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1053–1061 (2018)
63. Zhou, Z., Luo, Y., Wu, Y., Sun, K., Ji, J., Yan, K., Ding, S., Sun, X., Wu, Y., Ji, R.: Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. arXiv preprint arXiv:2507.02664 (2025)
64. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. Advances in Neural Information Processing Systems **36**, 77771–77782 (2023)