# CLAIMEVAL: ATOMIC-LEVEL EVALUATION OF EXPLANATIONS FOR AI-GENERATED IMAGE FORENSICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent work leverages MLLMs to build explainable AI-generated image detectors that output authenticity predictions and natural language explanations. However, these explanations are often unreliable—frequently generic, hallucinated, or forensically insufficient. Existing evaluation methods, including text similarity metrics, image-text similarity, and VLM-as-a-Judge, primarily produce holistic quality scores and cannot determine whether specific claims within an explanation are supported by visual evidence in the image, nor whether those claims constitute valid forensic evidence. We propose ClaimEval, a fine-grained evaluation framework that formulates explanation assessment as atomic-level forensic reasoning. Given an image-explanation pair, we decompose the explanation into atomic claims—the smallest independently verifiable semantic propositions—and evaluate each through a structured pipeline: language evaluation, claim grounding, evidence grouping, and forensic evidentiality assessment. We train a task-specific evaluator based on Qwen3-VL-8B with supervised fine-tuning followed by Structure-Guided GRPO (SG-GRPO). It produces structured claim-level judgments that are aggregated into an overall explanation score. To support this task, we construct a dataset with claim-level annotations and expert quality ratings. Experiments show that ClaimEval achieves substantially better alignment with human judgments than existing metrics. As a diagnostic tool, it reveals that classification accuracy does not imply explanation credibility, and serves as an effective inference-time reranker for improving detection reliability.

## 1 INTRODUCTION

The rapid advancement of photorealistic AI-generated images is blurring the boundary between synthetic and real content (Black Forest Labs, 2024; Stability AI, 2024), raising serious security and ethical concerns (Ferreira et al., 2020; Ren et al., 2024; Xu et al., 2023). Consequently, reliable AI-generated image detection has become increasingly important (Wang et al., 2020; Ojha et al., 2023; Yan et al., 2025b;a). Traditional methods formulate this task as binary classification without explanations of the underlying decision rationale (Ojha et al., 2023; Wang et al., 2020), which is insufficient in high-stakes scenarios where users need to understand *why* an image is judged as fake. Recent MLLM-based detectors output authenticity predictions and natural language explanations (He et al., 2025; Sun et al., 2025; Huang et al., 2025a; Xu et al., 2025; Zhou et al., 2025; Gao et al., 2026; Huang et al., 2026), but the mere availability of explanations does not ensure they are visually grounded or forensically meaningful.

Existing explainable detectors exhibit three recurring problems: generic statements without verifiable details about the relevant visual content (Li et al., 2023; Guan et al., 2024), hallucinated anomalies absent from the image (Rohrbach et al., 2018; Li et al., 2023; Guan et al., 2024), and observations that are visually present but forensically irrelevant. An explanation may correctly describe vivid colors or smooth skin, yet these properties also occur in real photographs and do not, by themselves, justify a judgment of AI generation or manipulation. Beyond linguistic soundness, evaluating such explanations therefore requires answering two distinct questions: **Grounding:** *Are the explanation's visual claims supported by the image?* **Evidentiality:** *Do the supported claims constitute valid forensic evidence?*



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

Existing holistic metrics do not explicitly distinguish these two questions. Text-generation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) measure agreement with reference texts without verifying image evidence. MLLMs can assess explanations as holistic judges (Chen et al., 2024), but an overall score does not reveal which claims lack visual support or which supported observations lack forensic significance.

We propose ClaimEval, a framework that evaluates forensic explanations through atomic claims: minimal, independently verifiable propositions. Claim grounding determines whether each visual observation is supported, refuted, or unresolved under the available image evidence. Because forensic significance may depend on relationships among observations, evidence grouping organizes related claims for joint interpretation. Forensic evidentiality assessment evaluates how strongly each resulting evidence unit—a group or an ungrouped eligible observation—supports an inference of AI generation or manipulation. Together with language evaluation, these judgments make explicit both the visual basis and the forensic significance of an explanation.

We train a Qwen3-VL-8B evaluator (Bai et al., 2025a) using supervised fine-tuning (SFT) followed by Structure-Guided GRPO (SG-GRPO) to produce all evaluation components jointly in a single structured output. Within this output, observations marked as unresolved because they require low-level image measurements are routed to tool-based verification, which updates their grounding status. The resulting structured judgments are then aggregated into an overall Explanation Score.

To support evaluator training and assessment, we construct a dataset of 6,150 image–explanation pairs from 2,160 images, covering general AI-generated images (Zhu et al., 2023), face forgery (Rössler et al., 2019), and image editing (Rombach et al., 2022). Explanations are collected from general-purpose MLLMs, specialized explainable detectors, and human annotators. Six domain experts review and refine model-assisted structured annotations for every sample across all data splits. Four experts additionally provide independent overall quality ratings for each test explanation, establishing a human reference for evaluating score alignment.

Our main contributions are:

- **A structured framework for forensic explanation evaluation.** We introduce ClaimEval to distinguish claim-level visual grounding from forensic evidentiality, with evidence grouping for joint interpretation and tool-based verification for low-level image properties. Its structured judgments support both error diagnosis and overall explanation scoring.
- **A human-reviewed dataset and a unified evaluator.** We construct 6,150 image–explanation pairs with expert-reviewed structured annotations across all data splits and independent expert quality ratings on the test set. Using these annotations, we train a unified evaluator through SFT and SG-GRPO to jointly produce the complete structured evaluation in a single output for each pair.
- **Evaluation with diagnostic and practical applications.** Experiments show stronger alignment with expert ratings than the evaluated baselines and reveal discrepancies between detection accuracy and explanation quality. ClaimEval also improves detection through Best-of-$N$ reranking, with cross-domain and reward-ablation analyses assessing generalization and component contributions.

## 2 Related Work

### 2.1 AI-Generated Image Detection and Explainable Forensics

AI-generated image detection seeks to distinguish real images from AI-generated or manipulated ones, covering generic AI-generated content (AIGC), DeepFake face manipulation, face swapping, and image editing. Existing methods investigate visual and statistical traces associated with image synthesis and manipulation, including spatial artifacts (Ojha et al., 2023; Yan et al., 2024; Tan et al., 2024a; NGUYEN et al., 2024; Fu et al., 2025; Yan et al., 2025a;b; Yang et al., 2026), frequency discrepancies (Qian et al., 2020; Tan et al., 2024b; Li et al., 2024; Kashiani et al., 2025), and, in video settings, temporal inconsistencies (Gu et al., 2021; 2022b;a; Yan et al., 2025c). To improve cross-domain generalization, studies explore spatial blending (Li et al., 2020a; Shiohara & Yamasaki, 2022; Zhao et al., 2021), frequency blending (Li et al., 2024; Kashiani et al., 2025), or feature-level augmentation (Yan et al., 2024). Datasets and benchmarks support evaluation across DeepFake and

general AIGC settings (Rössler et al., 2019; Dolhansky et al., 2019; Li et al., 2020b; Dolhansky et al., 2020; Zi et al., 2020; Zhou et al., 2021; Yan et al., 2025a; Zhang et al., 2025; Li et al., 2025; Huang et al., 2025b; Wang et al., 2025; Wen et al., 2025a; Huang et al., 2026). As MLLMs advance (Liu et al., 2023a; Bai et al., 2025b; Zhu et al., 2025), explainable detectors use language models to interpret vision detectors (Guo et al., 2025; Zhang et al., 2024; Huang et al., 2024; Peng et al., 2025) or employ LLM/MLLM reasoning for forensic analysis (He et al., 2025; Sun et al., 2025; Chen et al., 2026; Huang et al., 2025a; Xu et al., 2025; Zhou et al., 2025; Gao et al., 2026; Huang et al., 2026). These motivate evaluating explanation credibility alongside detection accuracy.

### 2.2 Evaluation of Explanations and Vision-Language Outputs

Explanation evaluation commonly draws on reference-based metrics that measure lexical overlap (Papineni et al., 2002; Lin, 2004; Banerjee & Lavie, 2005; Vedantam et al., 2015), semantic structure (Anderson et al., 2016), or contextual similarity (Zhang et al., 2020). CLIP-based measures assess global image–text semantic alignment (Radford et al., 2021; Hessel et al., 2021). LLM-as-a-Judge methods extend evaluation beyond reference similarity to criteria such as coherence, consistency, and relevance (Liu et al., 2023b; Zheng et al., 2023). This paradigm has been extended to vision-language tasks, where MLLMs score, compare, or rank candidate responses (Chen et al., 2024). However, multimodal judges can overlook relevant evidence or exhibit instability under semantically irrelevant perturbations (Lee et al., 2026). Complementary studies examine factual consistency and hallucinations in generated text and multimodal responses (Min et al., 2023; Jing et al., 2024; Cai et al., 2025). For forensic explanations, an overall quality judgment or factuality assessment alone does not establish whether visually supported observations constitute valid evidence of AI generation or manipulation. ClaimEval explicitly separates visual grounding from forensic evidentiality and evaluates related observations jointly as evidence groups, accounting for dependencies and redundant support. These structured judgments support both overall explanation scoring and claim-level diagnosis.

## 3 ClaimEval Evaluation Paradigm

ClaimEval evaluates forensic explanations through structured judgments about individual claims and their evidential relationships. The paradigm is supported by a claim-level annotated dataset, an evaluator trained with SFT and SG-GRPO, and tool-based verification for claims that require low-level image measurements.

### 3.1 Structured Evaluation Protocol

Given an image and its explanation, the ClaimEval evaluator jointly produces five complementary components in a single structured output, capturing textual validity, visual support, and forensic significance. **Language evaluation** identifies intrinsic textual errors, including repetition, semantic disconnection, and internal contradiction. **Atomic claim decomposition** represents the explanation as minimal, independently verifiable propositions, distinguishing visual observations from overall authenticity judgments. These claims provide the units for **claim grounding**, which assesses whether each observation is supported, contradicted, or unresolved under the available image evidence.

Because forensic significance can depend on relationships among observations, **evidence grouping** organizes related claims for joint interpretation, accounting for dependencies and redundant support. For example, an object's appearance and its reflection may be individually plausible yet reveal an inconsistency when considered together. Each group or ungrouped eligible observation forms a candidate evidence unit. **Forensic evidentiality assessment** evaluates the forensic significance of these units, distinguishing the presence of a visual property from its value as evidence of AI generation or manipulation. The complete annotation schema, label definitions, and eligibility constraints are provided in Appendix C.4.

### 3.2 Claim-Level Dataset

To support evaluator training and evaluation, we construct a claim-level dataset of 6,150 image–explanation pairs from 2,160 images, covering general AI-generated content, face manipulation,

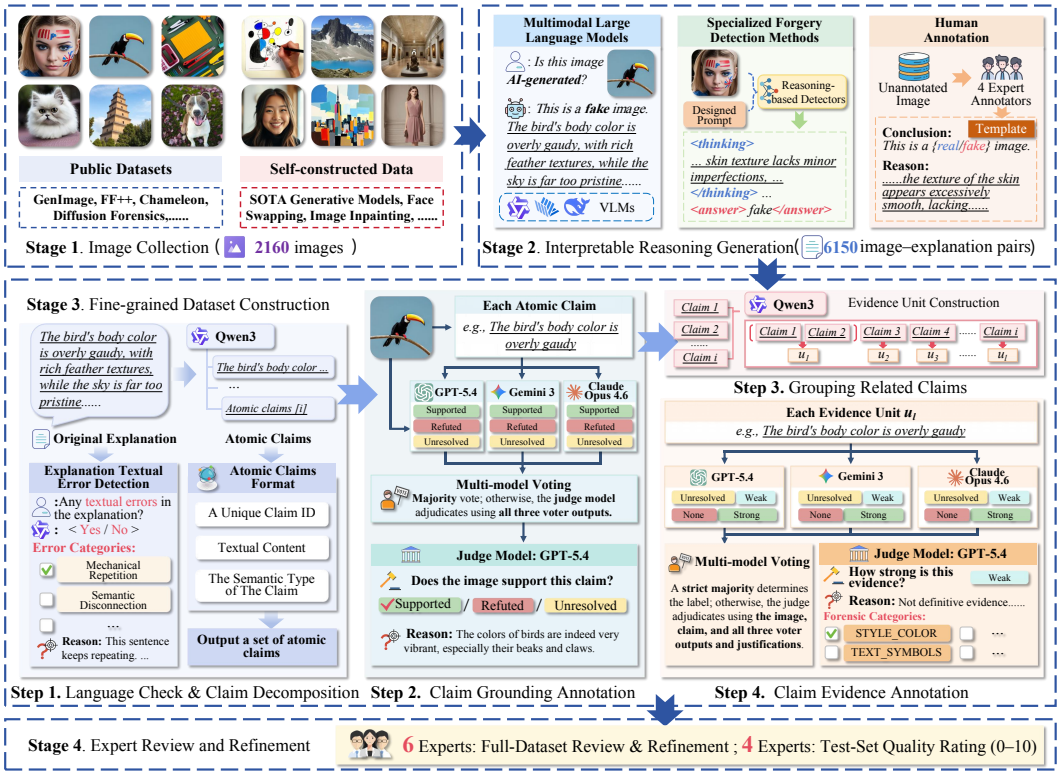



Figure 1: ClaimEval dataset construction. Images from public datasets and a self-constructed collection are explained by general-purpose MLLMs, specialized detectors, and annotators. Model-assisted annotations cover language evaluation, atomic claim decomposition, claim grounding, evidence grouping, and forensic evidentiality. Six domain experts review and refine annotations across all splits; four experts independently score test explanations (0–10) for human alignment evaluation.

and image editing. Figure 1 outlines the construction process, from image and explanation collection to structured annotation and expert refinement. Images are drawn from GenImage (Zhu et al., 2023), FF++ (Rössler et al., 2019), DiffusionForensics (Wang et al., 2023), and Chameleon (Yan et al., 2025a), supplemented with images produced by generative, face-swapping, and inpainting methods. For these images, we collect explanations from six general-purpose MLLMs, five specialized explainable detectors, and four expert annotators, capturing different explanation styles and approaches to forensic reasoning. We annotate these pairs following the structured evaluation protocol in Section 3.1. Qwen3-VL-235B first identifies textual errors and decomposes the explanation into atomic claims. For claim grounding, GPT-5.4, Gemini-3 Pro, and Claude Opus 4.6 independently assess each observation and provide a label and justification. A strict majority determines the label; otherwise, GPT-5.4 acts as the judge, using the image, claim, and all three judgments and justifications to resolve the case. Qwen3-VL-235B then organizes eligible observations into candidate evidence groups. Each resulting evidence unit, either a group or an ungrouped eligible observation, is assessed for forensic evidentiality using the same multi-model voting and adjudication procedure.

These automated annotations are candidates for expert review. Six domain experts in two teams of three review and refine the annotations across all data splits. Within each team, two experts independently review the annotations, and the third resolves disagreements. Their review covers language judgments, claim decomposition, grounding, evidence grouping, and forensic evidentiality. Four experts additionally assign independent overall quality scores on a 0–10 scale to the test explanations. The structured annotations support evaluator training and component-level assessment, while these overall ratings provide a human reference for evaluating the aggregated Explanation Score in Section 4.2. Detailed data splits and annotation statistics are provided in Appendix C.1, with the complete annotation schema in Appendix C.4 and human rating agreement in Appendix C.3.

### 3.3 TRAINING OF THE PROPOSED EVALUATOR

We train a unified Qwen3-VL-8B-Instruct evaluator in two stages. SFT learns complete structured evaluation from human-refined annotations, while Structure-Guided GRPO further optimizes claim fidelity, grounding accuracy, and evidence-structure agreement using task-specific rewards.

**Stage 1: SFT.** Let $\mathcal{D}_{\text{sft}}$ contain training triples $(\mathbf{x}_j, e_j, r_j^\star)$, where $\mathbf{x}_j$ is the image, $e_j$ is the explanation, and $r_j^\star$ is the structured reference annotation following Appendix C.4. Writing its token sequence as $s_j^\star = (s_{j,1}^\star, \ldots, s_{j,T_j}^\star)$, we train the evaluator policy $\pi_\theta$ by minimizing

$$\mathcal{L}_{\text{sft}} = -\mathbb{E}_{(\mathbf{x}_j, e_j, r_j^\star)\sim\mathcal{D}_{\text{sft}}} \frac{1}{T_j} \sum_{t=1}^{T_j} \log \pi_\theta \left( s_{j,t}^\star \mid \mathbf{x}_j, e_j, s_{j,<t}^\star \right), \tag{1}$$

where $T_j$ is the target length and $s_{j,<t}^\star$ denotes the preceding tokens. This objective supervises all five evaluation stages, including their labels, structural relationships, and textual justifications.

**Stage 2: SG-GRPO.** The resulting policy $\pi_{\text{sft}}$ initializes SG-GRPO and is retained as a frozen reference for KL regularization. Let $\mathcal{D}_{\text{sg}}$ denote the subset sampled from $\mathcal{D}_{\text{sft}}$ for SG-GRPO training. For each triple $(\mathbf{x}_j, e_j, r_j^\star) \in \mathcal{D}_{\text{sg}}$, the rollout policy $\pi_{\text{old}}$, a snapshot of the current policy, samples $G$ candidate outputs $\{\hat{r}_{j,g}\}_{g=1}^{G}$ conditioned on $(\mathbf{x}_j, e_j)$. Each candidate receives a reward $R_{j,g} = R(\hat{r}_{j,g}, r_j^\star)$ measuring agreement with the reference annotation. Let $\hat{s}_{j,g}$ denote its token sequence. The policy is optimized with

$$\begin{aligned}\mathcal{L}_{\text{SG-GRPO}} = -\mathbb{E}_{j,g,t}\Big[ \min \Big( \rho_{j,g,t}(\theta) A_{j,g}, \text{clip}\left(\rho_{j,g,t}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right) A_{j,g} \Big) \\ - \beta_{\text{KL}} D_{\text{KL}} \Big( \pi_\theta(\cdot \mid \mathbf{x}_j, e_j, \hat{s}_{j,g,<t}) \,\Big\|\, \pi_{\text{sft}}(\cdot \mid \mathbf{x}_j, e_j, \hat{s}_{j,g,<t}) \Big) \Big].\end{aligned} \tag{2}$$

Here, the expectation is over training examples, sampled candidates, and generated token positions. $\rho_{j,g,t}(\theta)$ is the token probability ratio between $\pi_\theta$ and $\pi_{\text{old}}$, and $A_{j,g}$ is the candidate's reward standardized within the $G$ outputs for the same input. $\epsilon_{\text{clip}}$ controls clipping, and $\beta_{\text{KL}}$ controls reference-policy regularization. The ratio and advantage definitions are given in Appendix B.2.

**Reward Design.** Each structured output is represented as $r = (\mathcal{L}, \mathcal{C}, \mathcal{G}, \mathcal{E}, \mathcal{U})$, corresponding to the five stages in Section 3.1. For a prediction $\hat{r}$ and reference $r^\star$, let $\chi(\hat{r}) \in \{0, 1\}$ indicate whether the prediction is JSON-parseable and satisfies the annotation schema. We define

$$R(\hat{r}, r^\star) = \begin{cases} -1, & \chi(\hat{r}) = 0, \\ \displaystyle\sum_{k\in\{\text{clm,grd,grp,evd}\}} \lambda_k R_k(\hat{r}, r^\star), & \chi(\hat{r}) = 1, \end{cases} \tag{3}$$

where $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$, and each component reward lies in $[0, 1]$.

Reward computation begins with type-constrained, one-to-one semantic matching between predicted and reference claims, using a frozen sentence encoder. The **atomic-claim fidelity reward** $R_{\text{clm}}$ combines claim coverage with matched semantic similarity. The **visual grounding reward** $R_{\text{grd}}$ credits grounding-status agreement on matched observations, while retaining missing reference observations and extraneous predicted observations in the normalization denominator. Building on the claim correspondence, the **evidence grouping reward** $R_{\text{grp}}$ measures agreement in group membership, group type, and relation type. The **forensic evidentiality reward** $R_{\text{evd}}$ compares evidentiality labels and forensic categories after matching evidence units by structure; category agreement contributes only when the evidentiality labels agree. The complete reward definitions, matching procedures, and hyperparameters are provided in Appendix B.3.

### 3.4 INFERENCE-TIME EVALUATION

During inference, observations labeled UNRESOLVED with reason type `TOOL_REQUIRED` are routed to a statistical verifier for properties such as noise, frequency characteristics, and compression traces. The verifier compares measurements of the claimed property with appropriate references,

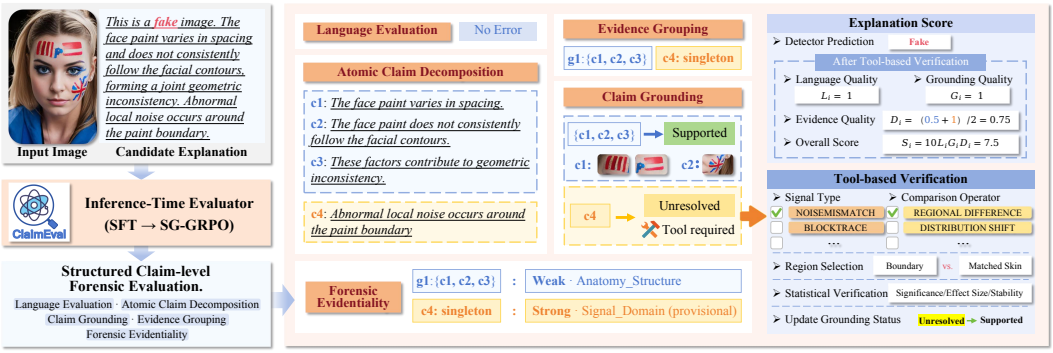



Figure 2: ClaimEval inference-time workflow. The evaluator jointly produces a structured five-component assessment per image–explanation pair. Tool verification updates grounding for `UNRESOLVED` observations marked `TOOL_REQUIRED`. Only units whose observations are all `SUPPORTED` after verification contribute to evidence quality. Language quality, updated grounding, and direction-aware evidence quality yield the Explanation Score.

updating grounding to SUPPORTED or REFUTED when a reliable decision is available and retaining UNRESOLVED otherwise. Tool-required observations participate provisionally in candidate evidence units. Group membership and evidentiality judgments remain unchanged; a unit contributes to evidence quality only if all its observations are SUPPORTED after verification.

The updated judgments are aggregated into an Explanation Score. For a schema-valid output on sample $i$, we compute

$$S_i = 10\, L_i G_i D_i, \qquad S_i \in [0, 10], \tag{4}$$

where $L_i$, $G_i$, and $D_i$ measure language quality, post-verification grounding quality, and evidence quality relative to the source detector's own prediction, respectively. The multiplicative form requires these dimensions to jointly support explanation credibility. Invalid outputs and outputs without valid evidence units receive zero. Original predictions and verification records are retained for diagnosis. Verification procedures and experiments are detailed in Appendix B.4, and the scoring factors are defined in Appendix B.5. Figure 2 summarizes the inference-time workflow.

## 4 EXPERIMENTS

We evaluate ClaimEval in five aspects: (1) fine-grained structured evaluation capability (Section 4.1), (2) alignment with human preferences (Section 4.2), (3) diagnostic benchmarking of explainable detectors (Section 4.3), (4) cross-domain generalization (Section 4.4), and (5) reward ablation (Section 4.5). We additionally demonstrate Best-of-$N$ reranking as a practical application (Section 4.6). Implementation details are provided in Appendix B.

### 4.1 FINE-GRAINED STRUCTURED EVALUATION

We first verify whether the evaluator learns the complete structured reasoning pipeline. To isolate the evaluator's learned structured-prediction capability, we evaluate the models' raw structured outputs against the original human-audited annotations, without tool-based verification. Table 1 reports eight metrics spanning all pipeline stages on the human-audited test set. Zero-shot MLLMs produce plausible language but struggle with structured forensic reasoning. Even the strongest zero-shot model, Claude Opus 4.6, achieves only 0.7582 Output Valid, 0.6574 Claim F1, 0.5489 Ground Acc., and 0.1013 Group F1@.75, with the particularly low grouping score indicating difficulty in organizing claims into coherent evidence units. Unified SFT nearly eliminates formatting failures, achieving 0.9988 Output Valid, and substantially improves claim decomposition (Claim F1 = 0.8785), grounding-status prediction (Ground Acc. = 0.7942), and evidence grouping (Group F1@.75 = 0.3629). SG-GRPO achieves perfect output validity (Output Valid = 1.0000) and further improves Claim F1 to 0.8873, Ground Acc. to 0.8276, and Group F1@.75 to 0.6286. The largest gain is observed in evidence grouping, with Group F1@.75 increasing by 0.2657 over Unified SFT.

Table 1: Fine-grained structured evaluation on the human-audited test set. Output Valid measures schema compliance; Lang. Acc. language-decision accuracy; Claim F1 and Sem. Sim. claim coverage and semantic fidelity; Ground Acc. grounding-status accuracy; Group F1@.75 evidence-group recovery at a similarity threshold of $0.75$; and Single Evid. Acc. and Group Evid. Acc. evidentiality accuracy for ungrouped and grouped evidence units, respectively. †: computed only on structurally matched evidence units. **Bold**: best; underline: second best.

| Method | Output Valid | Lang. Acc. | Claim F1 | Sem. Sim. | Ground Acc. | Group F1@.75 | Single Evid. Acc.† | Group Evid. Acc.† |
|---|---|---|---|---|---|---|---|---|
| *Zero-shot MLLMs* | | | | | | | | |
| Qwen3-VL-8B | 0.3817 | 0.8753 | 0.3984 | 0.8247 | 0.3862 | 0.0187 | 0.6245 | 0.4018 |
| Qwen3-VL-235B | 0.5269 | 0.9093 | 0.5197 | 0.8614 | 0.4478 | 0.0316 | 0.7284 | 0.5227 |
| GPT-5.4 | 0.6710 | 0.9321 | 0.6069 | 0.8918 | 0.4926 | 0.0607 | 0.8017 | 0.6039 |
| Gemini-3 Pro | 0.7160 | 0.9403 | 0.6295 | 0.9027 | 0.5158 | 0.0842 | 0.8249 | 0.6306 |
| Claude Opus 4.6 | 0.7582 | 0.9479 | 0.6574 | 0.9152 | 0.5489 | 0.1013 | 0.8461 | 0.6694 |
| gpt-5.6-luna | 0.6048 | 0.9461 | 0.6508 | 0.9136 | 0.4307 | 0.0488 | 0.8533 | 0.6481 |
| Unified SFT | 0.9988 | 0.9918 | 0.8785 | 0.9560 | 0.7942 | 0.3629 | 0.9388 | 0.9284 |
| SFT + SG-GRPO | **1.0000** | **0.9941** | **0.8873** | **0.9608** | **0.8276** | **0.6286** | **0.9462** | **0.9491** |

Table 2: Alignment with human-annotated explanation scores. Acc. measures agreement between the predicted and human three-level quality categories obtained by discretizing explanation scores into low, medium, and high levels based on score ranking and approximately equal-sized splits. Spearman $\rho$ and Kendall $\tau$ measure rank correlation between scores, while Pairwise Acc. measures agreement on pairwise quality comparisons. **Bold**: best; underline: second best.

| Method | Acc.↑ | Spearman $\rho$↑ | Kendall $\tau$↑ | Pairwise Acc.↑ |
|---|---|---|---|---|
| CLIPScore | 0.2336 | -0.0635 | -0.0261 | 0.4865 |
| Text Heuristic | 0.3665 | 0.2053 | 0.1359 | 0.5700 |
| Qwen3-VL-8B (Normal) | 0.3718 | 0.1478 | 0.0917 | 0.5538 |
| Qwen3-VL-235B (Normal) | 0.4561 | 0.2254 | 0.1734 | 0.6070 |
| GPT-5.4 (Normal) | 0.5211 | 0.2760 | 0.2124 | 0.6283 |
| Gemini-3 Pro (Normal) | 0.4608 | 0.2964 | 0.2218 | 0.6317 |
| gpt-5.6-luna (Normal) | 0.5023 | 0.2647 | 0.2076 | 0.6194 |
| Qwen3-VL-8B (Structured) | 0.4256 | 0.2315 | 0.1424 | 0.5742 |
| Qwen3-VL-235B (Structured) | 0.5111 | 0.2901 | 0.2474 | 0.6336 |
| GPT-5.4 (Structured) | 0.5667 | 0.3352 | 0.2654 | 0.6432 |
| Gemini-3 Pro (Structured) | 0.4783 | 0.3662 | 0.2908 | 0.6602 |
| gpt-5.6-luna (Structured) | 0.5585 | 0.3264 | 0.2576 | 0.6398 |
| **Ours** | **0.6212** | **0.5871** | **0.4254** | **0.7226** |

### 4.2 Alignment with Human Preference

We compare ClaimEval and baseline scores against the mean of four independent expert ratings (0–10) per test explanation ($\mathrm{ICC}(2,k) = 0.8824$; Appendix C.3). For ClaimEval, the aggregated Explanation Score is computed with the tool-augmented protocol: eligible UNRESOLVED claims with reason type `TOOL_REQUIRED` undergo tool-based verification before score aggregation. Other baselines use their respective scoring procedures. Table 2 compares our method against text heuristics, CLIPScore, and state-of-the-art MLLMs under two prompting settings: **Normal-prompt**, where the MLLM directly predicts a holistic quality score, and **Structured-prompt**, where each MLLM is prompted to follow the same structured evaluation protocol as ClaimEval.

Our evaluator consistently outperforms all baselines across both classification metrics (Accuracy) and ranking metrics (Spearman $\rho$, Kendall $\tau$, Pairwise Accuracy). Structured prompting improves over normal prompting for most MLLMs, confirming that claim-level decomposition helps even without task-specific training. However, even the best structured-prompt baseline (Gemini-3 Pro, $\rho$=0.3662) falls substantially behind our evaluator ($\rho$=0.5871), showing a substantial advantage for the full ClaimEval protocol over structured prompting alone.

Table 3: Classification and explanation-quality performance of explainable detectors on the General AIGC and DeepFake test sets. Acc., F1, Prec., and Recall measure real/fake classification performance, while Expl. Score denotes the aggregated explanation-quality score computed by ClaimEval. **Bold**: best result within each test set.

| Test Set | Method | Acc.↑ | F1↑ | Prec.↑ | Recall↑ | Expl. Score↑ |
|---|---|---|---|---|---|---|
| General AIGC | FakeVLM (Wen et al., 2025b) | **0.8318** | **0.8976** | 0.8259 | **0.9830** | 1.6738 |
| | SIDA (Huang et al., 2025a) | 0.6800 | 0.7459 | 0.9223 | 0.6261 | **2.4640** |
| | Legion (Kang et al., 2025) | 0.6658 | 0.7772 | 0.7772 | 0.7773 | 1.4799 |
| | FakeShield (Xu et al., 2025) | 0.6960 | 0.7619 | **0.9231** | 0.6487 | 1.2351 |
| DeepFake | FakeVLM (Wen et al., 2025b) | 0.7529 | 0.8482 | 0.7865 | 0.9204 | 1.2723 |
| | SIDA (Huang et al., 2025a) | 0.7652 | 0.8366 | 0.8750 | 0.8013 | 2.8134 |
| | Legion (Kang et al., 2025) | 0.7408 | 0.8424 | 0.7744 | 0.9234 | 1.2144 |
| | FakeShield (Xu et al., 2025) | 0.8195 | **0.8874** | 0.8339 | **0.9482** | 0.7953 |
| | FFAA (Huang et al., 2024) | 0.8086 | 0.8687 | **0.8947** | 0.8441 | 1.6815 |
| | Veritas (Tan et al., 2026) | **0.8288** | 0.8864 | 0.8827 | 0.8900 | **3.9617** |

Table 4: Cross-domain generalization. Metric definitions follow Table 1. **Bold**: better result within each setting.

| Setting | Method | Output Valid | Lang. Acc. | Claim F1 | Sem. Sim. | Ground Acc. | Group F1@.75 | Single Evid. Acc.† | Group Evid. Acc.† |
|---|---|---|---|---|---|---|---|---|---|
| Grouped-IID | Claude Opus 4.6 | 0.7582 | 0.9479 | 0.6574 | 0.9152 | 0.5489 | 0.1013 | 0.8461 | 0.6694 |
| | SFT + SG-GRPO | **1.0000** | **0.9941** | **0.8873** | **0.9608** | **0.8276** | **0.6286** | **0.9462** | **0.9491** |
| Image-OOD | Claude Opus 4.6 | 0.7415 | 0.9440 | 0.6390 | 0.9076 | 0.6220 | 0.1178 | 0.8300 | 0.6510 |
| | SFT + SG-GRPO | **0.9723** | **0.9858** | **0.8426** | **0.9448** | **0.7440** | **0.5853** | **0.9004** | **0.8872** |
| Expl.-OOD | Claude Opus 4.6 | 0.7485 | 0.9471 | 0.6468 | 0.9110 | **0.7362** | 0.0796 | 0.8395 | 0.6621 |
| | SFT + SG-GRPO | **0.9624** | **0.9815** | **0.8264** | **0.9387** | 0.7215 | **0.5410** | **0.8912** | **0.8695** |
| Joint-OOD | Claude Opus 4.6 | 0.7067 | 0.9245 | 0.5925 | **0.9088** | 0.5750 | 0.0849 | 0.7700 | 0.5880 |
| | SFT + SG-GRPO | **0.9200** | **0.9565** | **0.7521** | 0.9014 | **0.6128** | **0.5121** | **0.7805** | **0.7428** |

### 4.3 Benchmarking Existing Explainable Detectors

A fundamental premise of this paper is that classification accuracy does not imply explanation credibility. To demonstrate this, we apply ClaimEval as an independent diagnostic tool to benchmark recent explainable detectors on two large-scale test sets.

Table 3 shows frequent misalignment between classification performance and explanation quality. On the General AIGC set, FakeVLM achieves the highest accuracy, whereas SIDA obtains a higher Explanation Score despite its lower accuracy. FakeShield achieves the highest precision but the lowest Explanation Score. On DeepFake, however, Veritas achieves the highest accuracy and Explanation Score. Thus, classification metrics alone cannot fully characterize explanation quality.

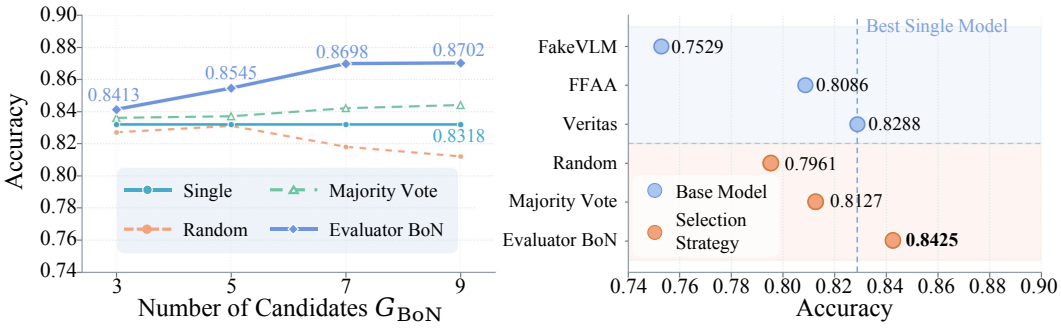
### 4.4 Cross-Domain Generalization

To assess robustness under controlled distribution shifts, we train separate ClaimEval evaluators in four training–test settings. **Grouped-IID** uses image domains and explanation sources observed during training, with no image overlap between training and test sets. For **Image-OOD**, we exclude all Chameleon images from training and evaluate on Chameleon, while retaining explanation sources observed during training. For **Explanation-OOD**, we exclude all Veritas-generated explanations from training and evaluate on Veritas explanations paired with images from in-distribution domains. Finally, **Joint-OOD** excludes both Chameleon images and Veritas-generated explanations from training and evaluates on Veritas explanations for Chameleon images. We compare each trained evaluator with Claude Opus 4.6, the strongest zero-shot baseline in Table 1.

As shown in Table 4, SFT + SG-GRPO consistently outperforms Claude Opus 4.6 across most distribution-shift settings and reported metrics. Under the most challenging Joint-OOD setting, it achieves 0.7521 Claim F1 and 0.5121 Group F1@.75, compared with 0.5925 and 0.0849 for Claude Opus 4.6, respectively. This result indicates that task-specific training improves robustness to simultaneous image-domain and explanation-source shifts.

Table 5: Reward ablation (higher is better). Metrics and notation follow Table 1.

| Variant | Output Valid | Lang. Acc. | Claim F1 | Sem. Sim. | Ground Acc. | Group F1@.75 | Single Evid. Acc.† | Group Evid. Acc.† |
|---|---|---|---|---|---|---|---|---|
| Unified SFT | 0.9988 | 0.9918 | 0.8785 | 0.9560 | 0.7942 | 0.3629 | 0.9388 | 0.9284 |
| w/o $R_{\text{clm}}$ | **1.0000** | 0.9936 | 0.8796 | 0.9567 | 0.8215 | 0.6069 | 0.9451 | 0.9476 |
| w/o $R_{\text{grd}}$ | **1.0000** | **0.9947** | 0.8880 | **0.9612** | 0.8004 | 0.5847 | 0.9367 | 0.9404 |
| w/o $R_{\text{grp}}$ | 0.9994 | 0.9924 | 0.8861 | 0.9599 | 0.8249 | 0.3518 | 0.9449 | 0.9470 |
| w/o $R_{\text{evd}}$ | **1.0000** | 0.9941 | **0.8882** | 0.9610 | **0.8292** | 0.6180 | 0.9342 | 0.9315 |
| SFT + SG-GRPO (Full) | **1.0000** | 0.9941 | 0.8873 | 0.9608 | 0.8276 | **0.6286** | **0.9462** | **0.9491** |




(a) Single-model multiple sampling

(b) Multi-model candidate pool

Figure 3: Best-of-$N$ reranking. (a) Selection among multiple FakeVLM outputs on General AIGC. (b) Selection from a heterogeneous pool of Veritas, FakeVLM, and FFAA outputs on DeepFake.

### 4.5 Reward Ablation

We remove each of the four reward components in turn and renormalize the remaining weights. Each reward primarily benefits its corresponding target (Table 5). Removing $R_{\text{grp}}$ causes the largest drop in Group F1@.75 (0.6286 $\rightarrow$ 0.3518), confirming that evidence reasoning supervision is critical for learning group structure. Removing $R_{\text{grd}}$ primarily affects Ground Acc. (0.8276 $\rightarrow$ 0.8004) and propagates to downstream evidence metrics. Removing $R_{\text{clm}}$ slightly reduces Claim F1 and Group F1. The full model delivers the strongest overall performance, particularly on evidence grouping and evidentiality evaluation, demonstrating that all four rewards contribute to evaluation quality.

### 4.6 Evaluator as a Test-Time Reranker (Best-of-N)

Beyond offline evaluation, we investigate whether the explanation score can improve downstream detection through Best-of-$N$ (BoN) reranking. In Fig. 3a, we sample $G_{\text{BoN}} \in \{3, 5, 7, 9\}$ outputs from FakeVLM and replace the default prediction only when a candidate exceeds it by a margin $\delta$ in explanation score. While random sampling degrades performance and majority vote quickly saturates, BoN improves accuracy from 83.18% to 87.02% over the single-pass baseline at $G_{\text{BoN}}$=9. In Fig. 3b, we construct a heterogeneous candidate pool from Veritas, FakeVLM, and FFAA and select the prediction with the highest explanation score. BoN achieves 0.8425 accuracy, outperforming majority vote (0.8127) and the best individual model (Veritas, 0.8288). These results demonstrate that ClaimEval generalizes across models and serves as an effective inference-time reranking module.

## 5 Conclusion

In this paper, we present ClaimEval, a fine-grained framework for evaluating explanations in AI-generated image detection. ClaimEval decomposes explanations into atomic claims and evaluates their linguistic validity, visual grounding, evidence grouping, and forensic evidentiality. We further construct a claim-level annotated dataset and train a task-specific evaluator with supervised fine-tuning followed by SG-GRPO. Experiments show that ClaimEval achieves stronger alignment with human judgments than existing metrics and general-purpose MLLMs, while also supporting robust cross-domain evaluation and effective inference-time reranking.

### AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic images, produce image-authenticity predictions and explanations, and assist with structured dataset annotation through claim decomposition, multi-model voting, and judge-based adjudication. Generative models were also used as explanation generators and evaluators in the experiments. Additionally, we used generative AI tools to polish the manuscript and correct grammatical errors. We reviewed all AI-assisted materials. In particular, six domain experts reviewed and refined the automated annotations across all dataset splits, resolving disagreements through adjudication, and four experts independently rated the quality of test-set explanations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

All authors have read and acknowledge the ICLR Code of Ethics and commit to adhering to its principles. We are committed to honest and transparent reporting, appropriate attribution of prior work, respect for data licenses and intellectual property, and protection of privacy and confidential information. These responsibilities apply to our use of facial imagery and our expert annotation process, including respect for annotators' rights and well-being. We recognize that dataset and annotation biases, assessment errors, and misuse may cause harm, and commit to disclosing relevant limitations and risks. ClaimEval is intended to support human review of forensic explanations; its scores should not serve as the sole basis for consequential judgments about individuals or image authenticity.

### REPRODUCIBILITY STATEMENT

Dataset construction and expert annotation are described in Section 3.2, with the data schema and benchmark composition provided in Appendix C. Section 3.3 and Appendix B describe the training objectives, hyperparameters, tool-based grounding verification, and explanation-scoring rules. The experimental settings are detailed in Section 4. We will publicly release our source code to support reproducibility.

## REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398. Springer, 2016.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Black Forest Labs. Announcing Black Forest Labs, 2024. URL https://bfl.ai/blog/24-08-01-bfl.

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Junzhe Chen, Xiaotao Gu, and Minlie Huang. MHALO: Evaluating MLLMs as fine-grained hallucination detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9197–9222. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.478. URL https://aclanthology.org/2025.findings-acl.478/.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24h.html.

Tao Chen, Jingyi Zhang, Decheng Liu, and Chunlei Peng. MGFFD-VLM: Multi-granularity prompt learning for face forgery detection with VLM. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 2707–2715, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792220. URL https://doi.org/10.1145/3774904.3792220.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset, 2019. URL https://arxiv.org/abs/1910.08854.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset, 2020. URL https://arxiv.org/abs/2006.07397.

William D. Ferreira, Cristiane B.R. Ferreira, Gelson da Cruz Júnior, and Fabrizzio Soares. A review of digital image forensics. *Computers & Electrical Engineering*, 85:106685, 2020. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2020.106685.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3040–3048, 2025.

Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Muxi Diao, Lei Chen, Kongming Liang, and Zhanyu Ma. Toward generalizable forgery detection and reasoning. *IEEE Transactions on Image Processing*, 35:3395–3410, 2026. doi: 10.1109/TIP.2026.3673967.

Google DeepMind. Gemini 3 Pro model card, 2025. URL https://deepmind.google/models/model-cards/gemini-3-pro/.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 3473–3481, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475508. URL https://doi.org/10.1145/3474085.3475508.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 744–752, 2022a.

Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European Conference on Computer Vision*, pp. 596–613. Springer, 2022b.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. doi: 10.1109/CVPR52733.2024.01363.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–116, 2025. doi: 10.1109/CVPR52734.2025.00019.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. VLForgery Face Triad: Detection, localization and attribution via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 163010–163044, 2025. doi: 10.52202/085713-5442. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ee94bf235482e4c1f689c04c81656dbf-Paper-Conference.pdf.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

OuCheng Huang, Manxi Lin, Jiexiang Tan, Xiaoxiong Du, Yang Qiu, Junjun Zheng, Xiangheng Kong, Yuning Jiang, and Bo Zheng. MIRAGE: Towards AI-generated image detection in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5076–5084, 2026.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, Wenming Yang, and Jiaya Jia. FFAA: Multimodal large language model based explainable open-world face forgery analysis assistant, 2024. URL https://arxiv.org/abs/2408.10072.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841, 2025a. doi: 10.1109/CVPR52734.2025.02685.

Zhenglin Huang, Xiangtai Li, Xi Yang, Bei Peng, Xiaowei Huang, Baoyuan Wu, Dacheng Tao, Ming-Hsuan Yang, and Guangliang Cheng. So-fake: Benchmarking and explaining social media image forgery detection, 2025b. URL https://arxiv.org/abs/2505.18660.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. LEGION: Learning to ground and explain for synthetic image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947, 2025. doi: 10.1109/ICCV51701.2025.01760.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8775–8785, 2025. doi: 10.1109/CVPR52734.2025.00820.

Sua Lee, Sanghee Park, and Jinbae Im. MM-JudgeBias: A benchmark for evaluating compositional biases in MLLM-as-a-Judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25336–25373. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1162. URL https://aclanthology.org/2026.acl-long.1162/.

Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. FreqBlender: Enhancing deepfake detection by blending frequency knowledge. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44965–44988, 2024. doi: 10.52202/079017-1429. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4f92d2f498b88f1bd43732312272967a-Paper-Conference.pdf.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5000–5009, 2020a. doi: 10.1109/CVPR42600.2020.00505.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213, 2020b. doi: 10.1109/CVPR42600.2020.00327.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-5739. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fb693c67f61e5321746ffce8b6fdd2d0-Paper-Datasets_and_Benchmarks_Track.pdf.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023a. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer Nature Switzerland, 2025. ISBN 978-3-031-72970-6.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Dat NGUYEN, Nesryne MEJRI, Inder Pal SINGH, Polina KULESHOVA, Marcella ASTRID, Anis KACEM, Enjie GHORBEL, and Djamila AOUADA. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17395–17405, 2024. doi: 10.1109/CVPR52733.2024.01647.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023. doi: 10.1109/CVPR52729.2023.02345.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. MLLM-enhanced face forgery detection: A vision-language fusion solution, 2025. URL https://arxiv.org/abs/2505.02013.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pp. 86–103. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Naina Raisinghani. Introducing Nano Banana Pro. Google Blog, November 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/. Accessed: 2026-09-20.

Jie Ren, Han Xu, Pengfei He, Yingqian Cui, Shenglai Zeng, Jiankun Zhang, Hongzhi Wen, Jiayuan Ding, Pei Huang, Lingjuan Lyu, Hui Liu, Yi Chang, and Jiliang Tang. Copyright protection in generative AI: A technical perspective, 2024. URL https://arxiv.org/abs/2402.02333.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019. doi: 10.1109/ICCV.2019.00009.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18699–18708, 2022. doi: 10.1109/CVPR52688.2022.01816.

Stability AI. Introducing Stable Diffusion 3.5, 2024. URL https://stability.ai/news-updates/introducing-stable-diffusion-3-5.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19576–19586, 2025. doi: 10.1109/CVPR52734.2025.01823.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024a. doi: 10.1109/CVPR52733.2024.02657.

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

Hao Tan, Jun Lan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 66420–66477, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6d13e085b79d454da5910e4ca82a3d9d-Paper-Conference.pdf.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.

Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. DFBench: Benchmarking deepfake image detection capability of large multimodal models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12666–12673, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758204. URL https://doi.org/10.1145/3746027.3758204.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8692–8701, 2020. doi: 10.1109/CVPR42600.2020.00872.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22388–22398, 2023. doi: 10.1109/ICCV51070.2023.02051.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. BusterX: MLLM-powered AI-generated video forgery detection and explanation, 2025a. URL https://arxiv.org/abs/2505.12620.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 58972–59005. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1974. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/551d275ae5cb1bb1ebe6e37fbe201db8-Paper-Conference.pdf.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 9291–9298, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612704. URL https://doi.org/10.1145/3581783.3612704.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 31186–31216, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4d4e0ab9d8ff180bf5b95c258842d16e-Paper-Conference.pdf.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b0303773962ea1b5394c3a83cc7dd066-Paper-Conference.pdf.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8984–8994, 2024. doi: 10.1109/CVPR52733.2024.00858.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/yan25b.html.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12615–12625, 2025c. doi: 10.1109/CVPR52734.2025.01177.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, and Xi Li. All Patches Matter, More Patches Better: Enhance AI-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, volume 2026, pp. 121742–121761, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c5ee2a08fbe743b171b0b4b2bdfd6f86-Paper-Conference.pdf.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer, 2026. URL https://arxiv.org/abs/2511.22699.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pp. 399–415. Springer, 2024.

Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-Bench: Are LMMs masters at evaluating AI-generated images? In *International Conference on Learning Representations*, volume 2025, pp. 84824–84845, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d355518527578ce26b80da96e9fc2750-Paper-Conference.pdf.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15003–15013, 2021. doi: 10.1109/ICCV48922.2021.01475.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5774–5784, 2021. doi: 10.1109/CVPR46437.2021.00572.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-Holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18746–18758, 2025. doi: 10.1109/ICCV51701.2025.01742.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

Mingjian Zhu, Hanting Chen, Qiangyu YAN, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4d4a021f9051a6c18183b059117e8b5-Paper-Datasets_and_Benchmarks.pdf.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390, 2020.

# APPENDIX

## A LIMITATIONS

Although ClaimEval provides a fine-grained and human-aligned framework for evaluating forensic explanations, it still has several limitations. First, our evaluator is built upon a specific MLLM backbone, and its performance may further benefit from stronger future multimodal models. Second, while our dataset covers diverse image sources, explanation sources, and forgery scenarios, it cannot exhaustively include all emerging generators and manipulation techniques. Third, the construction of structured claim-level annotations requires expert review, which introduces additional annotation cost. Nevertheless, the proposed framework is extensible: new data sources, stronger backbones, and additional forensic categories can be naturally incorporated to further improve its coverage and robustness.

## B IMPLEMENTATION DETAILS

### B.1 TRAINING SETUP

To ensure full reproducibility of the evaluator presented in the main text, this section outlines the detailed hyperparameter configurations and training procedures. All training experiments are conducted on 4 NVIDIA A100 GPUs (80GB). The training pipeline is divided into two sequential stages: (1) LoRA-based (Hu et al., 2022) SFT to instill the canonical structured evaluation format and basic claim-level reasoning, followed by (2) SG-GRPO to enforce structural stability and semantic claim-level alignment.

**Supervised Fine-Tuning.** In the initial stage, we fine-tune `Qwen3-VL-8B-Instruct`. We employ parameter-efficient tuning via LoRA, freezing the base model's backbone and updating only the LoRA parameters. The comprehensive configuration is detailed in Table 6.

Table 6: Supervised Fine-Tuning (SFT) configuration.

| Configuration | Value |
|---|---|
| Base model | `Qwen3-VL-8B-Instruct` |
| Fine-tuning method | LoRA |
| Batch size | 16 |
| Learning rate | $1 \times 10^{-4}$ |
| Epochs | 3 |
| Max image pixels | 401,408 |
| Max sequence length | 8,192 |
| LoRA rank / alpha / dropout | 8 / 32 / 0.05 |
| Target modules | `all-linear` |
| Optimizer | AdamW |
| Weight decay | 0.1 |
| LR scheduler | Cosine |
| Warmup ratio | 0 |
| Random seed | 42 |

**Structure-Guided GRPO (SG-GRPO).** To address practical generation artifacts—such as JSON formatting issues, suboptimal claim decomposition (over-splitting or under-splitting), or semantic deviations from human-reviewed reference claims—we further optimize the evaluator via SG-GRPO. The policy model is initialized from the SFT LoRA checkpoint, while the frozen SFT evaluator serves as the reference model for KL regularization to prevent the policy from diverging excessively. The RL training configuration is provided in Table 7.

**Reward and inference hyperparameters**. Table 8 summarizes the parameters used in reward computation and inference-time reranking.

Table 7: SG-GRPO training configuration.

| Configuration | Value |
|---|---|
| Samples per prompt ($G$) | 4 |
| Sampling temperature | 0.7 |
| Top-$p$ | 0.9 |
| Max completion length | 4,096 |
| Batch size | 8 |
| Learning rate | $5 \times 10^{-7}$ |
| Epochs | 1 |
| KL coefficient ($\beta_{\mathrm{KL}}$) | 0.04 |
| Clipping range ($\epsilon_{clip}$) | 0.2 |
| Reward normalization | Group-wise |
| Max image pixels | 401,408 |
| Max sequence length | 4,096 |
| Optimizer | AdamW |
| LR scheduler | Cosine |
| Random seed | 42 |

Table 8: Reward and inference hyperparameters used in ClaimEval. The similarity weights are listed in the same order as their corresponding definitions.

| Parameter description | Formal notation | Value |
|---|---|---|
| **Reward construction** | | |
| Claim-fidelity reward weight | $\lambda_{\mathrm{clm}}$ | 0.10 |
| Grounding reward weight | $\lambda_{\mathrm{grd}}$ | 0.40 |
| Evidence-grouping reward weight | $\lambda_{\mathrm{grp}}$ | 0.40 |
| Evidentiality reward weight | $\lambda_{\mathrm{evd}}$ | 0.10 |
| Claim-coverage coefficient | $\alpha_{\mathrm{c}}$ | 0.75 |
| Claim matching threshold | $\tau_c$ | 0.75 |
| Evidence-group similarity weights | $(\eta_{\mathrm{mem}}, \eta_{\mathrm{type}}, \eta_{\mathrm{rel}})$ | (0.50, 0.25, 0.25) |
| Evidence-unit similarity weights | $(\eta_{\mathrm{status}}, \eta_{\mathrm{cat}})$ | (0.80, 0.20) |
| Invalid-output reward | $R\|_{\chi=0}$ | $-1$ |
| **Best-of-$N$ inference** | | |
| Candidate counts for single-model sampling | $G_{\mathrm{BoN}}$ | $\{3, 5, 7, 9\}$ |
| Score margin for replacement | $\delta$ | 0.25 |

### B.2 SG-GRPO Optimization Details

For the sampled token $\hat{s}_{j,g,t}$, the probability ratio in Eq. (2) is

$$\rho_{j,g,t}(\theta) = \frac{\pi_\theta(\hat{s}_{j,g,t} \mid \mathbf{x}_j, e_j, \hat{s}_{j,g,<t})}{\pi_{\mathrm{old}}(\hat{s}_{j,g,t} \mid \mathbf{x}_j, e_j, \hat{s}_{j,g,<t})}. \tag{5}$$

The rollout policy $\pi_{\mathrm{old}}$ is held fixed during the corresponding policy update.

For the $G$ candidates sampled for the same input, the group-relative advantage is

$$A_{j,g} = \frac{R_{j,g} - \bar{R}_j}{\sigma_j + \epsilon_A}, \qquad \bar{R}_j = \frac{1}{G}\sum_{g=1}^{G} R_{j,g}, \tag{6}$$

where $R_{j,g} = R(\hat{r}_{j,g}, r_j^\star)$, $\sigma_j$ is the standard deviation of the $G$ rewards, and $\epsilon_A > 0$ is a numerical-stability constant. The SFT policy $\pi_{\mathrm{sft}}$ remains frozen throughout SG-GRPO as the reference for KL regularization.

### B.3 SG-GRPO Reward Details

This section defines the four component rewards in Eq. (3). For a predicted structured output $\hat{r}$ and its reference $r^\star$, hats and stars distinguish predicted and reference quantities, respectively. All

matching is performed within the same image–explanation pair, and $\mathbf{1}[\cdot]$ denotes the indicator function. The component rewards below apply to schema-valid predictions; invalid outputs receive the penalty specified in Eq. (3). Reward weights and similarity coefficients are reported in Table 8.

**Atomic-claim fidelity reward.** Let $\hat{\mathcal{C}} = \{\hat{c}_a\}_{a=1}^{\hat{n}_c}$ and $\mathcal{C}^\star = \{c_b^\star\}_{b=1}^{n_c^\star}$ denote the predicted and reference claim sets, where $\hat{n}_c$ and $n_c^\star$ are their respective sizes. A frozen sentence encoder measures the normalized semantic similarity between each pair of claim texts:

$$S_{ab} = \mathrm{sim}(\hat{c}_a, c_b^\star) \in [0, 1].$$

Let $\hat{y}_a^{\mathrm{type}}$ and $y_b^{\mathrm{type}\star}$ denote their claim types. We retain candidate pairs satisfying

$$S_{ab} \geq \tau_c, \qquad \hat{y}_a^{\mathrm{type}} = y_b^{\mathrm{type}\star},$$

where $\tau_c$ is the claim-matching threshold. Using $S_{ab}$ as the matching weight, we apply the Hungarian algorithm to obtain a maximum-weight one-to-one matching $\Omega_{\tau_c}$ over the retained pairs.

The claim reward combines coverage of the predicted and reference claims with the semantic fidelity of matched pairs:

$$R_{\mathrm{clm}}(\hat{r}, r^\star) = \alpha_c \frac{2|\Omega_{\tau_c}|}{\max(\hat{n}_c + n_c^\star, 1)} + (1 - \alpha_c) \frac{\sum_{(a,b)\in\Omega_{\tau_c}} S_{ab}}{\max(|\Omega_{\tau_c}|, 1)}, \tag{7}$$

where $\alpha_c \in [0, 1]$ balances claim coverage and matched semantic similarity. Unmatched claims receive no matching credit and remain in the coverage denominator. The correspondence $\Omega_{\tau_c}$ is also used to align observations and group members in the subsequent rewards.

**Visual grounding reward.** Grounding is evaluated on observation claims. We define the corresponding subset of matched claims as

$$\Omega_{\mathrm{obs}} = \left\{(a, b) \in \Omega_{\tau_c} : y_b^{\mathrm{type}\star} = \textsc{Observation}\right\}.$$

Because claim matching requires type agreement, both claims in each retained pair are observations. Let $\hat{n}_{\mathrm{obs}}$ and $n_{\mathrm{obs}}^\star$ denote the total numbers of predicted and reference observations, respectively, and let $\hat{y}_a^{\mathrm{grd}}$ and $y_b^{\mathrm{grd}\star}$ denote their grounding statuses. The grounding reward is

$$R_{\mathrm{grd}}(\hat{r}, r^\star) = \frac{2\sum_{(a,b)\in\Omega_{\mathrm{obs}}} \mathbf{1}\left[\hat{y}_a^{\mathrm{grd}} = y_b^{\mathrm{grd}\star}\right]}{\max(\hat{n}_{\mathrm{obs}} + n_{\mathrm{obs}}^\star, 1)}. \tag{8}$$

A matched observation receives credit only when its grounding status agrees with the reference. The denominator includes all predicted and reference observations, so missing and extraneous observations also reduce the reward.

**Evidence grouping reward.** Let $\hat{\mathcal{E}} = \{\hat{E}_p\}_{p=1}^{\hat{n}_\mathcal{E}}$ and $\mathcal{E}^\star = \{E_q^\star\}_{q=1}^{n_\mathcal{E}^\star}$ denote the predicted and reference evidence groups. Their respective numbers are $\hat{n}_\mathcal{E}$ and $n_\mathcal{E}^\star$. Because the groups contain claims from different index spaces, their members are compared through $\Omega_{\tau_c}$.

Let $\mathcal{I}(\hat{E}_p)$ and $\mathcal{I}(E_q^\star)$ denote the claim-index sets of the two groups. Their aligned member count is

$$M_{pq} = \sum_{(a,b)\in\Omega_{\tau_c}} \mathbf{1}[a \in \mathcal{I}(\hat{E}_p)]\mathbf{1}[b \in \mathcal{I}(E_q^\star)]. \tag{9}$$

The corresponding member overlap is

$$J_{pq} = \frac{M_{pq}}{|\mathcal{I}(\hat{E}_p)| + |\mathcal{I}(E_q^\star)| - M_{pq}}. \tag{10}$$

This matching-aware intersection-over-union accounts for both missing and extraneous members. In particular, unmatched predicted members remain in the denominator.

Let $\gamma$ and $v$ denote group type and relation type, respectively. The group-pair weight is

$$\begin{aligned} w_{pq}^{\mathcal{E}} &= \mathrm{sim}_{\mathcal{E}}(\hat{E}_p, E_q^\star) \\ &= J_{pq}\left(\eta_{\mathrm{mem}} + \eta_{\mathrm{type}}\mathbf{1}[\hat{\gamma}_p = \gamma_q^\star] + \eta_{\mathrm{rel}}\mathbf{1}[\hat{v}_p = v_q^\star]\right), \end{aligned} \tag{11}$$

where $\eta_{\text{mem}}, \eta_{\text{type}}, \eta_{\text{rel}} \geq 0$ and $\eta_{\text{mem}} + \eta_{\text{type}} + \eta_{\text{rel}} = 1$. Member overlap provides structural credit, with additional credit for agreement in group and relation types.

We exclude group pairs with $J_{pq} < \tau_{\mathcal{E}}$, where $\tau_{\mathcal{E}}$ is the minimum member-overlap threshold, and apply the Hungarian algorithm to the remaining pairs using $w^{\mathcal{E}}_{pq}$ as the matching weight. This yields a maximum-weight one-to-one matching $\Gamma$. The evidence grouping reward is

$$R_{\text{grp}}(\hat{r}, r^\star) = \frac{2\sum_{(p,q)\in\Gamma} w^{\mathcal{E}}_{pq}}{\max(\hat{n}_{\mathcal{E}} + n^\star_{\mathcal{E}}, 1)}. \tag{12}$$

Thus, the reward reflects both group recovery and the structural agreement of matched groups.

**Forensic evidentiality reward.** Let $\hat{\mathcal{U}} = \{\hat{u}_\ell\}_{\ell=1}^{\hat{n}_{\mathcal{U}}}$ and $\mathcal{U}^\star = \{u^\star_m\}_{m=1}^{n^\star_{\mathcal{U}}}$ denote the predicted and reference evidence-unit collections, with their associated evidentiality annotations. Here, $\hat{n}_{\mathcal{U}}$ and $n^\star_{\mathcal{U}}$ are the respective numbers of units. Each candidate unit is an evidence group or an ungrouped eligible observation, following Appendix C.4. Grouped claims are not evaluated again as singleton units.

We first establish unit correspondence using structural information. Identifying each unit with its underlying group or claim, its structural similarity is

$$s^{\text{str}}_{\ell m} = \begin{cases} J_{pq}, & \hat{u}_\ell = \hat{E}_p,\ u^\star_m = E^\star_q, \\ S_{ab}, & \hat{u}_\ell = \hat{c}_a,\ u^\star_m = c^\star_b,\ (a,b) \in \Omega_{\tau_c}, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Two group units are compared by aligned member overlap, while two singleton units are compared by matched claim similarity. Units of different structural types receive zero similarity.

We exclude pairs with $s^{\text{str}}_{\ell m} < \tau_{\mathcal{U}}$, where $\tau_{\mathcal{U}}$ is the unit structural-matching threshold, and apply the Hungarian algorithm to the remaining structural scores. The resulting maximum-weight one-to-one matching is denoted by $\Psi_{\text{str}}$. Evidentiality statuses and forensic categories do not affect this correspondence.

Let $\hat{y}^{\text{evd}}_\ell$ and $y^{\text{evd}\star}_m$ denote the evidentiality statuses of the matched units, and let $\hat{f}_\ell$ and $f^\star_m$ denote their forensic categories. The credit assigned to a unit pair is

$$\begin{aligned} w^{\mathcal{U}}_{\ell m} &= \text{sim}_{\mathcal{U}}(\hat{u}_\ell, u^\star_m) \\ &= s^{\text{str}}_{\ell m} \mathbf{1}\big[\hat{y}^{\text{evd}}_\ell = y^{\text{evd}\star}_m\big] \left(\eta_{\text{status}} + \eta_{\text{cat}} \mathbf{1}[\hat{f}_\ell = f^\star_m]\right), \end{aligned} \tag{14}$$

where $\eta_{\text{status}}, \eta_{\text{cat}} \geq 0$ and $\eta_{\text{status}} + \eta_{\text{cat}} = 1$. Evidentiality-status agreement acts as a hard gate: a status mismatch receives zero credit, and category agreement contributes only when the statuses match. For units labeled NONE or UNRESOLVED, the category is fixed to NOT_APPLICABLE by the annotation schema.

The forensic evidentiality reward is

$$R_{\text{evd}}(\hat{r}, r^\star) = \frac{2\sum_{(\ell,m)\in\Psi_{\text{str}}} w^{\mathcal{U}}_{\ell m}}{\max(\hat{n}_{\mathcal{U}} + n^\star_{\mathcal{U}}, 1)}. \tag{15}$$

This reward combines structural correspondence with evidentiality agreement, while accounting for missing and extraneous evidence units through normalization.

All four component rewards lie in $[0, 1]$. Empty matching sets contribute zero; under the definitions above, a component also receives zero when both its predicted and reference collections are empty. Language judgments and textual justifications are supervised through token-level SFT and are not separately scored by the four SG-GRPO reward components.

### B.4 TOOL-BASED VERIFICATION

The visual evaluator assigns the grounding status UNRESOLVED to an OBSERVATION claim when the image alone does not provide sufficient evidence for a reliable visual judgment. This commonly occurs for claims concerning low-level image properties, such as noise statistics, frequency

characteristics, neighborhood-pixel relations, and compression traces. If an OBSERVATION claim is labeled UNRESOLVED and assigned the reason type `TOOL_REQUIRED`, it is passed to the tool-based verification module described below.

### B.4.1 TOOL-BASED VERIFICATION PIPELINE

**Inputs and structured claim interpretation.** For an activated claim $c_i$, the module receives the image $\mathbf{x}$, the claim text $c_i$, and the associated grounding metadata

$$m_i = \left(\mathrm{id}_i,\, \phi_i^{\mathrm{grd}},\, q_i^{\mathrm{grd}}\right),$$

where $\mathrm{id}_i$ is the claim identifier, $\phi_i^{\mathrm{grd}}$ is the grounding reason type, and $q_i^{\mathrm{grd}}$ is the corresponding grounding justification.

The free-form claim and its metadata are converted into a structured measurement specification using `Qwen3-VL-8B-Instruct`. Specifically, the parser receives $(\mathbf{x}, c_i, m_i)$ and produces

$$s_i = \mathrm{Parse}_{\mathrm{Qwen}}\left(\mathbf{x}, c_i, m_i\right) = \left(\kappa_i,\, \varrho_i,\, \mathfrak{o}_i,\, a_i,\, \mathcal{S}_i,\, b_i\right),$$

where $\kappa_i$ is the signal family, $\varrho_i$ is the image representation, $\mathfrak{o}_i$ is the comparison direction or operator, $a_i$ is the target-region description, $\mathcal{S}_i$ is the analysis scope, and $b_i$ specifies the comparison basis. The comparison basis is either WITHINIMAGE, which uses comparable regions from the same image, or REALGALLERY, which uses a fixed gallery of authentic images. The parser also identifies the measurable property and the evidence required to verify it.

The parser is used only for semantic normalization. It does not calculate low-level statistics or make the final verification decision. If the claim cannot be mapped to a supported signal family, comparison operator, analysis scope, or comparison basis, the verification result is NOTIDENTIFIABLE.

**Signal types and verification directions.** Based on the low-level properties frequently invoked by existing MLLM-based explainable forgery detectors, we select the following representative signal families:

$$\mathcal{K} = \{\text{NOISELEVEL}, \text{NOISEMISMATCH}, \text{BANDSHAPE}, \text{PERIODICITY}, \text{BLOCKTRACE}\}.$$

These families cover the low-level property types considered in our experiments, including noise, frequency-spectrum shape, resampling-related periodicity, and localized compression traces. The set is not intended to be exhaustive. Additional signal families can be incorporated through the same verification interface by specifying their image representation, measurement function, comparison operator, and reference construction.

The corresponding image representations include the RGB image, a neighborhood-difference residual, a wavelet high-frequency representation, and an SRM residual.

The comparison operator $\mathfrak{o}_i$ specifies the direction of verification. It indicates whether the claim asserts that a statistic is larger than a reference value, smaller than a reference value, different from a reference distribution, or different between two spatial regions. The comparison basis $b_i$ determines whether the reference is constructed from comparable regions in the same image or from the real-image gallery.

**Spatial region extraction.** We perform statistical verification on rectangular image regions or patches. If a claim specifies explicit coordinates or a canonical image-relative position, the corresponding rectangular region is constructed directly. For object- or part-specific descriptions, we use Grounding DINO (Liu et al., 2025), which takes the image $\mathbf{x}$ and target description $a_i$ as inputs and predicts bounding boxes for the described content.

The selected box defines the target region in the original image coordinates. We represent this region by a binary indicator $\mathbf{m}_i$, which equals one inside the rectangle and zero outside. Statistical measurements are computed over the selected region or patches within it. Grounding DINO provides spatial localization; the verification decision is determined by the subsequent statistical analysis.

When $b_i$ = WITHINIMAGE, candidate control patches are sampled outside the target rectangle. Each patch $p$ is represented by

$$\mathbf{h}(p) = \left(\mu_{\mathrm{lum}}(p),\, \sigma_{\mathrm{tex}}(p),\, \sigma^2_{\mathrm{grad}}(p)\right),$$

where $\mu_{\mathrm{lum}}$ denotes mean luminance, $\sigma_{\mathrm{tex}}$ denotes local texture variation, and $\sigma^2_{\mathrm{grad}}$ denotes gradient variance. Let $\mathbf{h}_i^{\mathrm{target}}$ denote the corresponding descriptor of the target region. A candidate control patch is retained when

$$\left\|\mathbf{h}(p) - \mathbf{h}_i^{\mathrm{target}}\right\|_2 \leq \tau_{\mathrm{match}},$$

where $\tau_{\mathrm{match}}$ is the control-patch matching threshold. This matching reduces the influence of natural differences between regions with substantially different texture or structure.

When $b_i = \textsc{RealGallery}$, the same measurement family, image representation, comparison operator, and applicable spatial scope are applied to the images in the fixed real-image gallery $\mathcal{D}_{\mathrm{real}}$. The resulting measurements form the reference set for claims that explicitly compare the target property with authentic images.

If the target region cannot be localized, is too small for reliable analysis, or does not have a sufficient number of comparable control patches or reference measurements, the verification result is set to NotIdentifiable.

**Image-condition checks.** Before computing the target statistic, the module evaluates image conditions that may affect low-level measurements, including compression artifacts, high-frequency preservation, global noise level, and effective spatial resolution. These quantities are used as quality gates and are not treated as evidence of manipulation.

For example, frequency-periodicity claims are not evaluated when the image is dominated by compression artifacts or has insufficient effective resolution. Similarly, absolute noise-level claims are not evaluated when strong denoising makes the original noise scale unidentifiable. A failed quality gate results in NotIdentifiable rather than Inconsistent.

**Low-level measurements.** For a selected signal family $\kappa_i$, the module computes a deterministic measurement

$$\widehat{\theta}_i = \mathcal{M}_{\kappa_i}\left(\mathbf{x}, \mathbf{m}_i, \varrho_i\right),$$

where $\mathcal{M}_{\kappa_i}$ denotes the measurement function associated with signal family $\kappa_i$.

For noise-related claims, the luminance image is denoted by $y$. We compute a four-neighbor residual

$$\xi(\mathbf{p}) = y(\mathbf{p}) - \frac{1}{4}\left(y(\mathbf{p}_{\mathrm{up}}) + y(\mathbf{p}_{\mathrm{down}}) + y(\mathbf{p}_{\mathrm{left}}) + y(\mathbf{p}_{\mathrm{right}})\right),$$

where the four subscripts indicate the vertically and horizontally adjacent pixels. The robust noise scale is estimated as

$$\widehat{\sigma}_{\mathrm{noise}} = \frac{\mathrm{median}\left(|\xi - \mathrm{median}(\xi)|\right)}{0.6745}.$$

For a regional claim, the target estimate is compared with the estimates from the matched control patches.

For frequency-band claims, let $P(q)$ denote the mean power at radial frequency $q$. The spectral shape is summarized by the slope of the log–log radial power spectrum:

$$\widehat{\beta} = \mathrm{slope}\left(\log(q+1),\, \log(P(q) + \epsilon_{\mathrm{num}})\right),$$

where $\epsilon_{\mathrm{num}} > 0$ is a numerical-stability constant. Periodicity claims are evaluated from the prominence of peaks in the residual power spectrum. JPEG block claims are evaluated by comparing discontinuities along the $8 \times 8$ block grid with local discontinuities away from the grid. Neighborhood-pixel claims use the corresponding neighborhood-difference representation.

**Reference comparison and verification decision.** The reference measurements depend on the comparison basis $b_i$. Let $\mathcal{V}_i^{\mathrm{ref}}$ denote the reference measurements:

$$\mathcal{V}_i^{\mathrm{ref}} = \begin{cases} \left\{\widehat{\theta}_{i,k}^{\mathrm{ctrl}}\right\}_{k=1}^{K_i}, & b_i = \textsc{WithinImage}, \\ \left\{\widehat{\theta}_{i,k}^{\mathrm{real}} : \mathbf{x}_{i,k}^{\mathrm{real}} \in \mathcal{D}_{\mathrm{real}}\right\}, & b_i = \textsc{RealGallery}. \end{cases}$$

Here, $\widehat{\theta}_{i,k}^{\mathrm{ctrl}}$ denotes the measurement from the $k$-th retained control patch, $\widehat{\theta}_{i,k}^{\mathrm{real}}$ denotes the measurement from the $k$-th selected real reference image or region, and $K_i$ is the number of retained reference measurements. The symbol $\mathbf{x}_{i,k}^{\mathrm{real}}$ denotes the $k$-th selected image from the real-image gallery.

For either comparison basis, the reference statistic and its dispersion are computed as

$$\widehat{\theta}_i^{\mathrm{ref}} = \frac{1}{K_i} \sum_{k=1}^{K_i} \widehat{\theta}_{i,k}^{\mathrm{ref}}, \qquad \widehat{\sigma}_i^{\mathrm{ref}} = \mathrm{Std}\left(\mathcal{V}_i^{\mathrm{ref}}\right),$$

where $\mathrm{Std}(\cdot)$ denotes the standard deviation.

Let $\widehat{\theta}_i^{\mathrm{target}}$ denote the statistic computed on the target region. The raw effect and standardized effect are

$$\widehat{\Delta}_i = \widehat{\theta}_i^{\mathrm{target}} - \widehat{\theta}_i^{\mathrm{ref}}, \qquad d_i = \frac{\widehat{\Delta}_i}{\widehat{\sigma}_i^{\mathrm{ref}} + \epsilon_{\mathrm{num}}}.$$

The verification record also stores a confidence interval

$$\mathrm{CI}_i = \left[\widehat{\Delta}_i^{-}, \widehat{\Delta}_i^{+}\right],$$

where $\widehat{\Delta}_i^{-}$ and $\widehat{\Delta}_i^{+}$ are the lower and upper confidence limits, respectively.

Let $z_i^{\mathrm{tool}}$ denote the final tool-based verification decision. For directional claims, the adjusted hypothesis test and the minimum practical effect requirement are evaluated jointly:

$$z_i^{\mathrm{tool}} = \begin{cases} \textsc{Consistent}, & p_i^{\mathrm{adj}} \le \alpha_{\mathrm{sig}} \ \wedge \ |d_i| \ge \Delta_{\mathrm{min}} \ \wedge \ \mathrm{sign}(d_i) \text{ agrees with } \mathfrak{o}_i, \\ \textsc{Inconsistent}, & p_i^{\mathrm{adj}} \le \alpha_{\mathrm{sig}} \ \wedge \ |d_i| \ge \Delta_{\mathrm{min}} \ \wedge \ \mathrm{sign}(d_i) \text{ contradicts } \mathfrak{o}_i, \\ \textsc{NotIdentifiable}, & \text{otherwise.} \end{cases}$$

Here, $p_i^{\mathrm{adj}}$ is the adjusted $p$-value, $\alpha_{\mathrm{sig}}$ is the statistical significance level, $\Delta_{\mathrm{min}}$ is the minimum practical effect threshold, and $\mathrm{sign}(\cdot)$ denotes the sign function.

For claims that assert a difference without specifying a direction, a two-sided comparison is used. Stability is evaluated under small perturbations to the target-box position and size, patch size, control-patch selection, and FFT or wavelet window placement. The resulting measurements and verification decisions are compared across perturbations. These records are used together with the image-condition checks to determine whether the result is identifiable. Finally, the tool decision is mapped to the grounding status as follows:

$$\begin{aligned} \textsc{Consistent} &\rightarrow \textsc{Supported}, \\ \textsc{Inconsistent} &\rightarrow \textsc{Refuted}, \\ \textsc{NotIdentifiable} &\rightarrow \textsc{Unresolved}. \end{aligned}$$

For score aggregation, the mapped grounding statuses determine which candidate evidence units are valid. Verification does not reconstruct evidence groups or revise evidentiality judgments. A candidate group is retained only if all its members are SUPPORTED; its members are not reassigned as singleton units when the group is excluded. The original evaluator output and tool-verification records are retained separately for diagnosis.

### B.4.2 Empirical Evaluation of Tool-Based Grounding Verification

**Dataset construction.** We construct an intervention-based benchmark from real images to evaluate tool-based grounding verification. The benchmark contains both object-level and region-level interventions. Object-level interventions are applied to a selected object or object part, whereas region-level interventions are applied to a localized image region. Localized interventions are applied within rectangular regions specified by image coordinates or localized from object descriptions using Grounding DINO.

The benchmark includes seven intervention settings: regional noise addition, regional denoising, regional noise mismatch, regional sharpening, regional low-pass filtering, regional JPEG recompression, and global bicubic upsampling. For each setting with an explicit intensity parameter, the intervention strength is randomly selected from a predefined three-level set during data construction. In particular, the levels are $\sigma \in \{10, 20, 35\}$ for noise addition, strength $\in \{5, 10, 20\}$ for denoising, $\alpha_{\text{sharp}} \in \{1.5, 2.5, 4.0\}$ for sharpening, and cutoff $\in \{0.3, 0.5, 0.7\}$ for low-pass filtering. The selected level is sampled independently for each intervention instance. Settings without an explicit intensity parameter use their corresponding fixed operation configuration.

For every intervention instance, an unmodified image with the same source content is retained as a matched control. We generate 1000 intervention instances for each of the seven settings, resulting in 7000 intervention instances in total. Each intervention instance is paired with one unmodified control, yielding 7000 control instances and 14000 image instances overall. The intervention samples are constructed so that the expected verification decision is CONSISTENT, while the matched unmodified controls are used to estimate false-alarm behavior.

The real-image gallery $\mathcal{D}_{\text{real}}$ is used when a claim requires comparison with authentic images. In the intervention benchmark, the paired unmodified controls provide the primary reference for evaluating the corresponding intervention.

**Evaluation metrics.** Let $\mathcal{I}^+$ denote the intervention samples and $\mathcal{I}^0$ denote the unmodified control samples. The consistency detection rate is

$$\text{DR} = \frac{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} = \textsc{Consistent}\right]}{|\mathcal{I}^+|}.$$

The abstention rate is

$$\text{AR} = \frac{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} = \textsc{NotIdentifiable}\right]}{|\mathcal{I}^+|}.$$

The decision coverage is

$$\text{Coverage} = 1 - \text{AR} = \frac{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} \neq \textsc{NotIdentifiable}\right]}{|\mathcal{I}^+|}.$$

Among intervention samples receiving a decisive result, selective precision is

$$\text{SP} = \frac{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} = \textsc{Consistent}\right]}{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} \neq \textsc{NotIdentifiable}\right]}.$$

The corresponding selective error rate is

$$\text{SE} = \frac{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} = \textsc{Inconsistent}\right]}{\sum_{v \in \mathcal{I}^+} \mathbf{1}\left[z_v^{\text{tool}} \neq \textsc{NotIdentifiable}\right]} = 1 - \text{SP}\,.$$

Finally, the control false-alarm rate is

$$\text{FAR} = \frac{\sum_{v \in \mathcal{I}^0} \mathbf{1}\left[z_v^{\text{tool}} = \textsc{Consistent}\right]}{|\mathcal{I}^0|}.$$

Selective precision and selective error are reported together with coverage, because high selective precision can otherwise be obtained by abstaining on most samples.

**Results.** The highest coverage is obtained for regional noise mismatch, reaching 86.9%. Denoising and bicubic upsampling achieve the next highest coverage, with 84.3% and 84.1%, respectively. Noise addition yields 82.7% coverage, while sharpening, regional JPEG recompression, and low-pass filtering obtain 76.2%, 72.5%, and 70.7%, respectively. All intervention settings except regional JPEG recompression achieve 100.0% selective precision among decisive samples. Regional JPEG recompression is more difficult, obtaining 72.5% coverage and 99.3% selective precision, with a

Table 9: Evaluation of tool-based grounding verification on the intervention-based benchmark. Coverage denotes the fraction of intervention samples receiving a decisive result. Selective precision and selective error are computed only among decisive intervention samples, whereas FAR is computed over all unmodified control samples.

| **Intervention** | $N$ | **Coverage** | **Sel. Prec.** | **Sel. Error** | **FAR** |
|---|---|---|---|---|---|
| Add noise | 1000 | 82.7% | 100.0% | 0.0% | 2.0% |
| Denoise | 1000 | 84.3% | 100.0% | 0.0% | 4.0% |
| Regional noise mismatch | 1000 | 86.9% | 100.0% | 0.0% | 2.0% |
| Sharpen | 1000 | 76.2% | 100.0% | 0.0% | 6.0% |
| Low-pass filter | 1000 | 70.7% | 100.0% | 0.0% | 8.0% |
| Regional JPEG recompress | 1000 | 72.5% | 99.3% | 0.7% | 3.0% |
| Bicubic upsample | 1000 | 84.1% | 100.0% | 0.0% | 0.0% |
| **Overall** | **7000** | **79.6%** | **99.9%** | **0.1%** | **3.6%** |

0.7% selective error rate. Because the source images already contain compression traces, the additional local recompression signal is more difficult to distinguish from existing regional variation. Across all 7000 intervention samples, the verifier obtains 79.6% decision coverage and 99.9% selective precision among decisive samples. Its selective error rate is 0.1%. The false-alarm rate over the 7000 matched unmodified control instances is 3.6%.

### B.5 Overall Explanation Score

The overall Explanation Score evaluates whether an explanation provides linguistically valid, visually grounded, and forensically meaningful support for the detector's own prediction. It is used for benchmarking explanation quality and is distinct from the reward functions used to train SG-GRPO.

For sample $i$, let

$$r_i = (\mathcal{L}_i, \mathcal{C}_i, \mathcal{G}_i, \mathcal{E}_i, \mathcal{U}_i)$$

denote the structured output of ClaimEval. Prediction hats are omitted in this section for clarity. We use $y^{\text{type}}(c)$ and $y^{\text{grd}}(c)$ to denote the type and grounding status of claim $c$, and $y^{\text{evd}}(u)$ to denote the evidentiality status of evidence unit $u$.

**Valid evidence units.** Let

$$\mathcal{O}_i = \left\{ c \in \mathcal{C}_i : y^{\text{type}}(c) = \textsc{Observation} \right\}$$

denote the set of observation claims. A valid final evidence unit is either (1) an evidence group whose members are all SUPPORTED observations, or (2) an ungrouped SUPPORTED observation treated as a singleton. FORENSIC_CONCLUSION claims are excluded, and a grouped claim cannot also appear as a singleton. Let $\mathcal{U}_i^{\text{val}} \subseteq \mathcal{U}_i$ denote the resulting valid evidence units. Each group is treated as one unit, preventing related or repeated claims from receiving redundant credit.

All quantities below are computed after any `TOOL_REQUIRED` claims have been processed by the statistical verification module and their grounding status has been updated.

**Language quality.** The language factor $L_i$ is determined from the error categories recorded in $\mathcal{L}_i$ using fixed category-based weights:

$$L_i = \begin{cases} 1.0, & \text{no error,} \\ 0.8, & \text{mechanical repetition or other text errors,} \\ 0.5, & \text{semantic disconnection or incompleteness,} \\ 0, & \text{internal contradiction, gibberish, or formatting breakdown.} \end{cases} \tag{16}$$

When multiple error categories are present, $L_i$ takes the lowest applicable value.

**Grounding quality.** For the observation claims in $\mathcal{O}_i$, define

$$n_{\text{sup}}^{(i)} = \sum_{c \in \mathcal{O}_i} \mathbf{1}\left[y^{\text{grd}}(c) = \textsc{Supported}\right],$$
$$n_{\text{ref}}^{(i)} = \sum_{c \in \mathcal{O}_i} \mathbf{1}\left[y^{\text{grd}}(c) = \textsc{Refuted}\right],$$
$$n_{\text{unres}}^{(i)} = \sum_{c \in \mathcal{O}_i} \mathbf{1}\left[y^{\text{grd}}(c) = \textsc{Unresolved}\right], \tag{17}$$

and

$$N_{\text{obs}}^{(i)} = n_{\text{sup}}^{(i)} + n_{\text{ref}}^{(i)} + n_{\text{unres}}^{(i)}.$$

The grounding factor is

$$G_i = \begin{cases} \dfrac{n_{\text{sup}}^{(i)}}{N_{\text{obs}}^{(i)}} \max\left(0, 1 - \dfrac{n_{\text{ref}}^{(i)} + 0.5 n_{\text{unres}}^{(i)}}{N_{\text{obs}}^{(i)}}\right), & N_{\text{obs}}^{(i)} > 0, \\ 0, & N_{\text{obs}}^{(i)} = 0. \end{cases} \tag{18}$$

Thus, REFUTED claims receive a full penalty, whereas unresolved grounding receives a half penalty and no positive evidence credit.

**Direction-aware evidence quality.** Let $\hat{d}_i \in \{\texttt{FAKE}, \texttt{REAL}\}$ denote the authenticity prediction supplied by the source detector that generated explanation $e_i$. For a valid evidence unit $u \in \mathcal{U}_i^{\text{val}}$, let $\phi^{\text{evd}}(u)$ denote its forensic reason type. When the detector predicts FAKE, the directional credit is

$$w_{\text{FAKE}}(u) = \begin{cases} 1.0, & y^{\text{evd}}(u) = \textsc{Strong}, \\ 0.5, & y^{\text{evd}}(u) = \textsc{Weak}, \\ 0, & y^{\text{evd}}(u) \in \{\textsc{None}, \textsc{Unresolved}\}. \end{cases} \tag{19}$$

When the detector predicts REAL, the directional credit is

$$w_{\text{REAL}}(u) = \begin{cases} 1.0, & y^{\text{evd}}(u) = \textsc{None} \\ & \text{and } \phi^{\text{evd}}(u) = \texttt{NATURALLY_EXPLAINABLE}, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

The direction-aware evidence factor is

$$D_i = \begin{cases} \dfrac{1}{|\mathcal{U}_i^{\text{val}}|} \displaystyle\sum_{u \in \mathcal{U}_i^{\text{val}}} w_{\hat{d}_i}(u), & |\mathcal{U}_i^{\text{val}}| > 0, \\ 0, & |\mathcal{U}_i^{\text{val}}| = 0. \end{cases} \tag{21}$$

All valid evidence units remain in the denominator, including units with zero directional credit.

**Output validity.** Outputs that cannot be parsed or do not satisfy the required schema receive an Explanation Score of zero. An output with no valid evidence unit also receives zero because $D_i = 0$.

**Sample-level score.** The factors defined above are combined using Eq. (4). Schema-invalid outputs receive zero, as do outputs without valid evidence units. The score uses no fitted saturation or calibration parameters.

# C Dataset Details

## C.1 Dataset Composition and Statistics

The annotated dataset contains 6,150 image–explanation pairs from 2,160 images. The SFT split comprises 4,442 explanations for 1,560 images, while the test split contains 1,708 explanations for 600 images. SG-GRPO uses a subset of the SFT split consisting of 1,139 explanations for 400

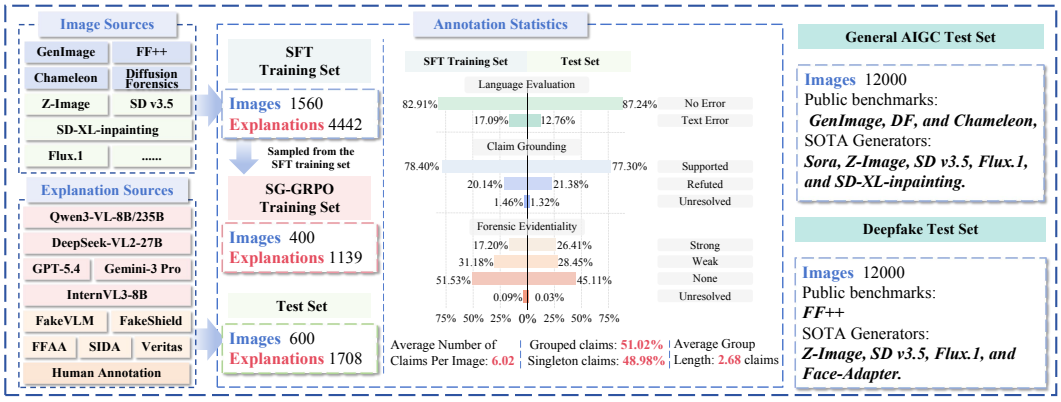



Figure 4: Dataset composition and annotation statistics, including image and explanation sources, training and test splits, label distributions, and evidence-group statistics. The separate General AIGC and DeepFake diagnostic benchmarks are also summarized.

images. Figure 4 summarizes the data sources, split sizes, annotation distributions, and evidence-group statistics.

The General AIGC and DeepFake diagnostic benchmarks each contain 12,000 images and do not overlap with the 2,160 images in the annotated dataset. Their source composition is reported in Table 12.

### C.2 Data Collection and Annotation Details

Images are collected from two sources: public datasets and a self-constructed collection. The public datasets include GenImage (Zhu et al., 2023), FF++ (Rössler et al., 2019), DiffusionForensics (Wang et al., 2023), and Chameleon (Yan et al., 2025a). The self-constructed collection contains images produced using Stable Diffusion V3.5 Large (Stability AI, 2024), Z-Image (Z-Image Team et al., 2026), Gemini Nano Banana Pro (Raisinghani, 2025), and face-swapping and image-inpainting methods.

The six general-purpose explanation generators are Qwen3-VL-8B and Qwen3-VL-235B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), DeepSeek-VL2-27B (Wu et al., 2024), GPT-5.4 (OpenAI, 2026), and Gemini-3 Pro (Google DeepMind, 2025). The five specialized detectors are FakeVLM (Wen et al., 2025b), Veritas (Tan et al., 2026), FFAA (Huang et al., 2024), SIDA (Huang et al., 2025a), and FakeShield (Xu et al., 2025). Human-written explanations follow a three-step template: describe the visual content, explain why the observations are problematic or reasonable, and provide an overall real/fake judgment.

Structured annotation and expert review follow the procedure in Section 3.2. For the independent test-set quality ratings, experts consult standardized visualizations as supplementary evidence when an explanation invokes low-level image properties. These overall ratings are used to assess alignment with human judgments.

### C.3 Human Rating Reliability for Explainable Content

This section assesses agreement among the four experts who independently rated test-set explanation quality.

To validate the reliability of the overall human ratings for explainable content, we conduct a consistency analysis on the evaluation results from 4 annotators. Recognizing the inherent subjectivity in explanation evaluation, our primary metric is relative ranking consistency rather than absolute score parity. As demonstrated in Table 10, pairwise agreement is robust, yielding an average Pearson $r$ of 0.7599. Furthermore, Table 11 reports an ICC(2,k) of 0.8824 and a Krippendorff's alpha of 0.6514. These results support the use of averaged expert ratings for evaluating alignment with human judgments.

Table 10: Pairwise agreement between human annotators for explanation-quality ratings.

| **Annotator Pair** | **Pearson** $r \uparrow$ | **Spearman** $\rho \uparrow$ | **Kendall** $\tau \uparrow$ |
|---|---|---|---|
| A1 vs. A2 | 0.7821 | 0.7626 | 0.5726 |
| A1 vs. A3 | 0.7446 | 0.7390 | 0.5385 |
| A1 vs. A4 | 0.8024 | 0.7893 | 0.6102 |
| A2 vs. A3 | 0.7148 | 0.7134 | 0.5187 |
| A2 vs. A4 | 0.7628 | 0.7581 | 0.5644 |
| A3 vs. A4 | 0.7527 | 0.7413 | 0.5186 |
| **Average** | **0.7599** | **0.7506** | **0.5538** |

Table 11: Overall reliability metrics for human explanation-quality scores.

| **Metric** | **Value** |
|---|---|
| ICC(2,k) | 0.8824 |
| Krippendorff's alpha | 0.6514 |

### C.4 Dataset Format

Each dataset record is represented as a multimodal instruction–response pair. The record stores the input image $\mathbf{x}$, the explanation $e$ and the corresponding structured annotation.

The evaluator receives only the image–explanation pair $(\mathbf{x}, e)$ and returns a single structured JSON object $r = (\mathcal{L}, \mathcal{C}, \mathcal{G}, \mathcal{E}, \mathcal{U})$. All records use the same field names, controlled labels, identifier conventions, and cross-field constraints.

The response is organized as a logically ordered but unified structured record. The following subsections describe the role of each component, its fields and admissible values, and the structural constraints that connect it to the other components.

#### C.4.1 Language Evaluation

The field `caption_language_evaluation` describes intrinsic textual quality independently of image content.

**`has_error`:** a Boolean value indicating whether the explanation contains a textual error.

**`error_types`:** a list of one or more controlled error labels.

**`explanation`:** a concise explanation of the detected textual problem. For an explanation with no error, the field briefly states why the text is considered valid.

The controlled textual-error labels are:

- `MECHANICAL_REPETITION`;
- `SEMANTIC_DISCONNECTION_INCOMPLETENESS`;
- `GIBBERISH_FORMATTING_BREAKDOWN`;
- `INTERNAL_CONTRADICTION`;
- `OTHER_TEXT_ERROR`;
- `NO_ERROR`.

Fig. 5 visualizes these categories, highlighting the exact spans that trigger specific error labels.

#### C.4.2 Atomic Claim Decomposition

The field `atomic_claims` contains the minimal independently verifiable propositions extracted from the explanation.

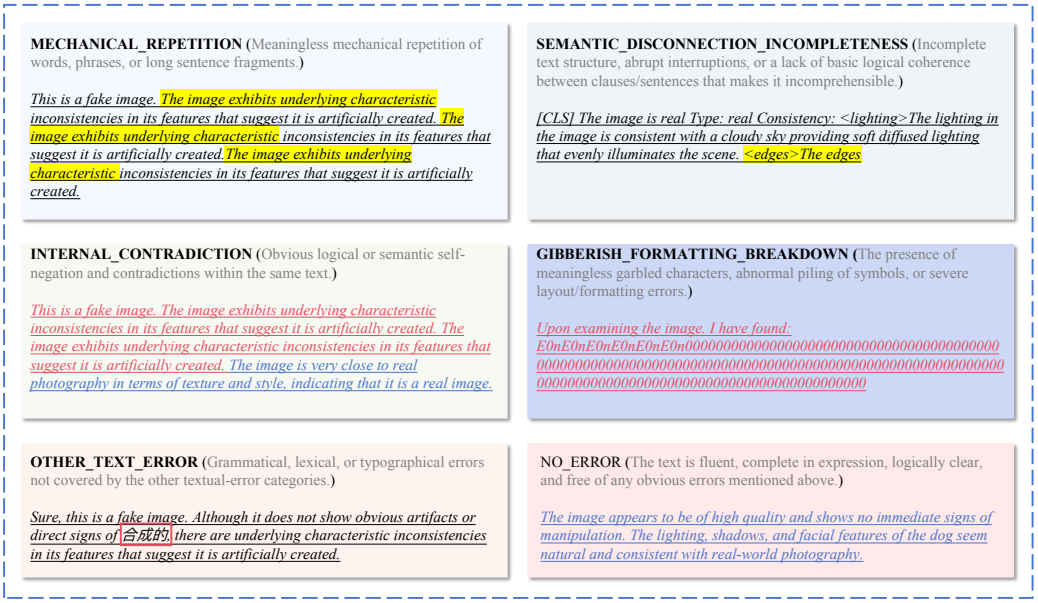



Figure 5: Taxonomy of textual validity labels. Representative examples for the six textual validity categories utilized in our evaluation protocol. Highlighted text denotes the specific evidence triggering the error label.

**`claim_id`:** a unique identifier for each claim within one sample.

**`text`:** the textual content of the atomic claim.

**`claim_type`:** the semantic type of the claim.

The claim-type options are:

- `OBSERVATION`: a visual proposition that can be checked against the image, including direct visual facts and relational visual statements;
- `FORENSIC_CONCLUSION`: an overall authenticity judgment about whether the image is real, edited, manipulated, or AI-generated.

### C.4.3 Claim Grounding

The field `claim_grounding` evaluates whether each `OBSERVATION` claim is supported by the image. Grounding records are assigned only to observation claims. Overall authenticity conclusions are excluded from claim grounding and evidence-unit construction.

**`claim_id`:** the identifier of the corresponding observation claim.

**`status`:** the three-way grounding judgment.

**`reason_type`:** the controlled reason for the grounding judgment.

**`required_evidence`:** nullable auxiliary information describing the evidence required for verification. It is `null` unless the reason type is `REFERENCE_REQUIRED` or `TOOL_REQUIRED`.

**`reason`:** a concise explanation supporting the grounding decision.

The grounding-status options are:

- `SUPPORTED`: the image provides sufficient visible evidence for the claim;
- `REFUTED`: the image clearly contradicts the claim;

- UNRESOLVED: the claim cannot be determined from the available image evidence because of occlusion, low resolution, ambiguity, missing context, tool requirements, or non-visual content.

The grounding reason vocabulary includes the following types:

- DIRECTLY_VISIBLE: the claim describes directly observable image content;
- EXPLICIT_CONTRADICTION: the image explicitly contradicts the claim;
- OCCLUDED: the relevant content is partially or fully occluded;
- LOW_RESOLUTION: image quality is insufficient for reliable verification;
- REFERENCE_REQUIRED: verification requires an external reference unavailable from the image alone;
- TOOL_REQUIRED: verification requires statistical, signal, frequency, noise, or pixel-level analysis;
- NON_VISUAL: the claim concerns a property that cannot be verified by visual inspection alone.

UNRESOLVED claims with reason type TOOL_REQUIRED are eligible for tool-based verification. The required_evidence field specifies the measurement or reference needed to perform this verification.

#### C.4.4 EVIDENCE GROUP CONSTRUCTION

The field evidence_groups identifies candidate observation claims that should be interpreted jointly.

**group_id:** a unique identifier for each evidence group.

**claim_ids:** the identifiers of the claims belonging to the group.

**group_type:** the structural type of the relation among the group members.

**relation_type:** the semantic relation that explains why the claims should be interpreted jointly.

**combination_reason:** a concise explanation of the joint evidential role of the group.

The group-type options are:

- EXPLICIT_RELATION: one OBSERVATION explicitly describes the relation and other eligible OBSERVATION claims provide its visual components;
- IMPLICIT_RELATION: multiple eligible OBSERVATION claims form a meaningful relation although no single claim explicitly states it.

The relation-type options are:

- JOINT_INCONSISTENCY: multiple observations jointly reveal an inconsistency;
- COMPARATIVE_DEVIATION: the evidence depends on comparison between parts, regions, objects, or expected visual patterns;
- COMPOSITIONAL_DEPENDENCY: the interpretation of one claim depends on another claim in the same scene;
- REDUNDANT_SUPPORT: multiple observations provide mutually reinforcing support for the same forensic interpretation.

A claim belongs to at most one group. SUPPORTED claims and UNRESOLVED claims with reason type TOOL_REQUIRED may participate in candidate evidence groups. Groups containing unresolved tool-required claims remain provisional until tool-based verification is completed. Claims with REFUTED grounding status and unresolved claims whose reason type is not TOOL_REQUIRED cannot participate in evidence-group construction.

### C.4.5 FORENSIC EVIDENTIALITY

The field `forensic_evidentiality` assigns forensic significance to the candidate evidence units. Each record corresponds either to a candidate group or to an ungrouped eligible observation claim.

**`evidence_unit_id`:** the identifier of the underlying evidence group or claim, such as `g1` or `c4`.

**`unit_type`:** either `EVIDENCE_GROUP` or `SINGLE_CLAIM`.

**`status`:** the forensic significance of the evidence unit.

**`reason_type`:** the controlled reason for the assigned evidentiality status.

**`forensic_category`:** the type of forensic phenomenon associated with the unit.

**`reason`:** a concise explanation of the assigned evidentiality status and category.

The evidence-status options are:

- `STRONG`: the unit provides compelling evidence supporting AI generation or manipulation;
- `WEAK`: the unit describes an ambiguous or non-diagnostic anomaly that may also occur naturally or result from compression;
- `NONE`: the unit provides no affirmative evidence of AI generation or manipulation;
- `UNRESOLVED`: the forensic significance cannot be determined from the available evidence.

The forensic reason-type options are:

- `CONCLUSIVE_FORENSIC_ANOMALY`: the unit identifies a specific, highly diagnostic anomaly that provides strong evidence of AI generation or manipulation;
- `AMBIGUOUS_FORENSIC_ANOMALY`: the unit identifies a suspicious anomaly that may also result from natural variation, compression, or other benign processes;
- `GENERIC_OR_VAGUE`: the unit makes a forensic judgment without identifying a sufficiently specific, verifiable property;
- `NATURALLY_EXPLAINABLE`: the unit addresses a specific forensic concern through an image-supported benign explanation or a concrete physical or geometric consistency check relevant to that concern;
- `IRRELEVANT_TO_FORENSICS`: the unit describes scene content without providing a rationale relevant to authenticity assessment;
- `INSUFFICIENT_FORENSIC_SUPPORT`: the available observations do not adequately support the proposed forensic interpretation;
- `FORENSIC_ASSESSMENT_UNAVAILABLE`: the unit's forensic significance cannot be reliably assessed under the available conditions;
- `REFERENCE_REQUIRED`: assessing the unit's forensic significance requires an external reference that is not available from the image alone;

The forensic-category options are:

- `ANATOMY_STRUCTURE`: anatomical, geometric, or structural inconsistencies in people or objects;
- `TEXT_SYMBOLS`: visual, spatial, or semantic inconsistencies in text, symbols, or logos;
- `LIGHTING_PHYSICS`: inconsistencies in illumination, reflections, perspective, depth, or other physical cues;
- `TEXTURE_RENDERING`: irregularities in texture, material appearance, detail, or local smoothness;
- `UNKNOWN_OBJECT`: implausible object presence, identity, or part structure;
- `STYLE_COLOR`: unnatural or inconsistent color, tone, saturation, style, or overall appearance;

- `SIGNAL_DOMAIN`: anomalies in low-level signal statistics, noise patterns, frequency characteristics, or pixel relations;
- `NOT_APPLICABLE`: used when the evidence status is `NONE` or `UNRESOLVED`, and for generic `WEAK` evidence when no specific category is justified.

For an evidence unit with `unit_type` equal to `EVIDENCE_GROUP`, `evidence_unit_id` refers to the corresponding `group_id`. For an evidence unit with `unit_type` equal to `SINGLE_CLAIM`, `evidence_unit_id` refers to the corresponding `claim_id`. Grouped claims are evaluated jointly and are not evaluated again as individual evidence units. Each candidate group and each ungrouped eligible observation has one evidentiality record. Eligible observations are SUPPORTED claims or UNRESOLVED claims with reason type `TOOL_REQUIRED`. The latter receive provisional evidentiality assessments; a candidate unit contributes to evidence-quality scoring only when all its observations are SUPPORTED after verification.

Fig. 6 pairs representative images with atomic claims corresponding to these categories, emphasizing that each claim serves as an independently verifiable semantic unit.

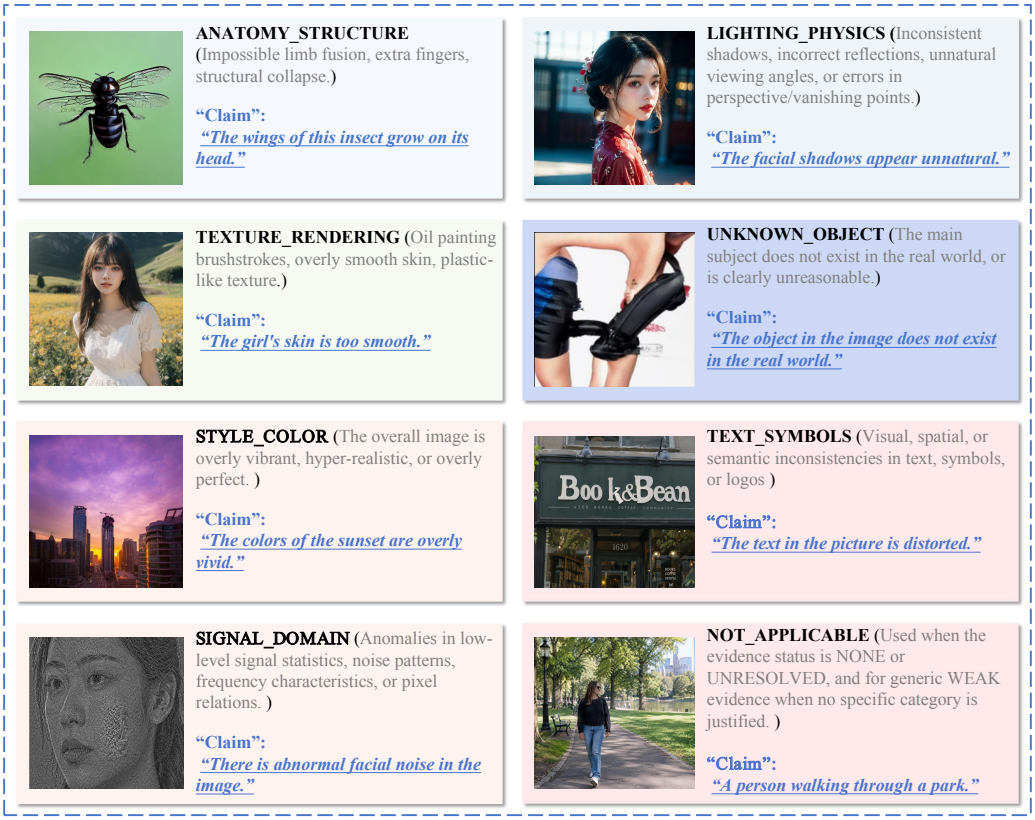



Figure 6: Claim-level forensic anomaly categories. Representative visual phenomena paired with their atomic claims, demonstrating the diverse axes of forensic evidence evaluated by our system.

### C.5 DIAGNOSTIC BENCHMARK COMPOSITION

To comprehensively evaluate explainable forgery detection methods across diverse real-world scenarios, our test sets incorporate images sourced from a wide range of origins:

- **GenImage**: A comprehensive large-scale benchmark for AI-generated image detection covering multiple generators. Authentic real-world images for our General AIGC test set are also sampled from its real image subset.
- **DiffusionForensics**: A specialized forensic dataset focusing specifically on identifying artifacts and traces left by state-of-the-art diffusion models.

Table 12: Detailed composition of the large-scale image test sets. To evaluate broad generalization, our test sets incorporate existing benchmarks, advanced generative/manipulation methods, and authentic real-world data.

| Test Set | Specific Generator / Source | # Images |
|---|---|---|
| General AIGC Image Test Set | GenImage | 1,000 |
| | DiffusionForensics | 1,000 |
| | Chameleon | 2,000 |
| | Sora | 1,000 |
| | Z-Image | 1,000 |
| | SD v3.5 | 1,000 |
| | Flux.1 | 1,000 |
| | SD-XL-inpainting | 1,000 |
| | Real Images (Sampled from GenImage) | 3,000 |
| | **Total** | **12,000** |
| DeepFake Image Test Set | FF++ | 1,000 |
| | Z-Image | 2,000 |
| | SD v3.5 | 2,000 |
| | Flux.1 | 2,000 |
| | Face-Adapter | 2,000 |
| | Real Images (Sampled from FFHQ) | 3,000 |
| | **Total** | **12,000** |

- **Chameleon**: A challenging evaluation dataset designed to test the robustness of detection models against highly photorealistic and diverse AI-generated images.
- **FaceForensics++ (FF++)**: A widely adopted, large-scale dataset explicitly constructed for facial manipulation detection, featuring various face-swapping and reenactment techniques.
- **Sora, SD v3.5, Flux.1, Z-Image**: The latest state-of-the-art text-to-image and text-to-video generative models, capable of producing exceptionally detailed and photorealistic synthetic visuals.
- **SD-XL-1.0-inpainting**: An advanced diffusion-based inpainting tool utilized to generate locally manipulated images, testing the models' fine-grained anomaly detection capabilities.
- **Face-Adapter**: Sophisticated techniques and deep learning frameworks specialized for high-fidelity facial identity swapping and precise editing.
- **FFHQ (Flickr-Faces-HQ)**: A high-quality image dataset of human faces. We exclusively sample authentic facial images from FFHQ to serve as the real baseline in our DeepFake test set.

Detailed compositions of the two 12,000-image test sets (General AIGC and DeepFake) are provided in Table 12. The two sets of 12,000 diagnostic test images do not overlap with the 2,160 images used for training and testing.