CLOVER: COVERAGE-LED OPTIMIZATION OF VERIFIABLE EVIDENCE REASONING FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors must recognize forgeries across diverse generators and scenes and provide explanations. Binary labels are inexpensive to obtain but provide no supervision for explanations, whereas evidence annotations are costly and may still omit valid observations. To this end, we introduce CLOVER, a training framework that expands detection coverage through binary supervision and advances explanation learning through open-ended evidence verification. For coverage-led supervised fine-tuning, we combine a large binary-labeled image collection for detection with a smaller, scene-diverse evidence corpus for explanation learning. This allows new forgery sources to enter training at low annotation cost with only binary labels. Atomic evidence reinforcement learning then improves evidence quality in two stages: first encouraging the discovery of more visually supported observations, then strengthening penalties for unsupported or uncertain claims. During this process, we employ a self-judge to assess the model's open-ended evidence against the image one claim at a time, enabling rewards that require authenticity labels but no reference explanations. Experiments show that supervised training already achieves strong detection and explanation performance with an evidence corpus roughly 10% the size of the binary corpus. Subsequent evidence optimization further improves precision and reduces hallucinations, demonstrating that visual verification can improve explanation reliability without collecting additional reference explanations.

1 INTRODUCTION

Advances in image generation and the misuse of synthetic content make reliable AI-generated image detection increasingly important. Detectors must recognize forgeries across diverse generators and scenes and explain their decisions with verifiable visual evidence (Zhu et al., 2023; Xia et al., 2025; Tan et al., 2026). A promising approach is to adapt multimodal large language models (MLLMs) for this task (Xu et al., 2025; Huang et al., 2025a; Wen et al., 2026).

However, training such detectors poses two supervision challenges. First, binary labels are inexpensive to obtain but provide no supervision for explanations. Requiring evidence annotations for every source would make broad detection coverage costly. Second, reference explanations may omit valid observations, and a correct real-or-fake decision does not guarantee that its explanation is supported by the image (Tan et al., 2026). These challenges raise two questions: **(1)** How can detection coverage expand without detailed annotations for every source? **(2)** How can explanation quality improve without complete reference explanations?

To address the first question, we introduce CLOVER, a coverage-led training framework. Detection benefits from exposure to diverse forgery sources, while explanation learning requires examples of what to observe in different scenes. Cues such as malformed text or inconsistent object geometry can recur across generators. We therefore assign broad source coverage to inexpensive binary labels and concentrate evidence annotations on scene diversity. Joint supervised fine-tuning (SFT) connects both capabilities in a shared model, allowing new sources to enter training without collecting explanations for each source.

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

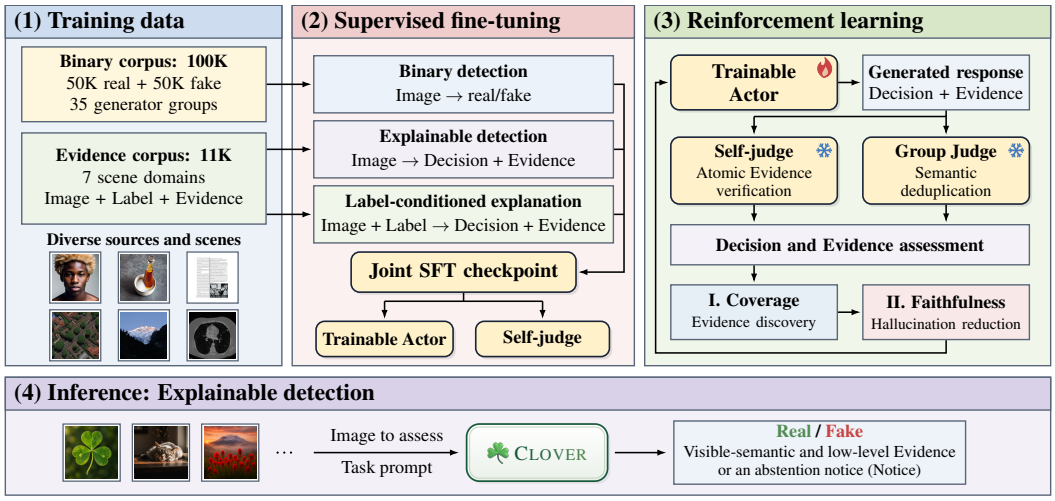


Figure 1: **CLOVER separates source coverage from evidence learning.** Broad binary supervision supports detection, while scene-diverse evidence annotations train explainable detection and label-conditioned explanation. Two-stage reinforcement learning improves evidence quality without reference explanations. Notice denotes evidence abstention. Flames mark parameter updates; snowflakes mark frozen parameters. Benchmark thumbnails are from GenImage (Zhu et al., 2023), LOKI (Ye et al., 2025), and So-Fake (Huang et al., 2025b).

For the second question, an observation can be valid even when it is absent from a reference explanation. We therefore employ a self-judge to assess open-ended claims against the image and use its judgments with authenticity labels to guide atomic evidence reinforcement learning (RL). Training first encourages evidence discovery and then strengthens penalties for unsupported or uncertain claims. This order gives the model more evidence to draw on before it learns to be selective.

Experiments show that detection can expand to new sources with a fixed evidence corpus. Further evidence optimization improves precision and reduces hallucinations while largely preserving detection performance.

Our main contributions are as follows:

1. **Coverage-led training framework.** We introduce CLOVER, which decomposes explainable detection into three complementary tasks jointly optimized during supervised fine-tuning and reinforcement learning. Binary detection covers diverse forgery sources, while explanation learning emphasizes scene coverage. Adding new forgery sources to training therefore requires only inexpensive binary labels.
2. **Open-ended atomic evidence reinforcement learning.** We use visual verification to improve explanation reliability without reference explanations. Rewards encourage visually supported evidence and count repeated observations only once. Two-stage optimization first encourages evidence discovery and then strengthens penalties for unsupported or uncertain claims.
3. **Performance with limited evidence supervision.** CLOVER achieves the best average macro-F1 across five detection benchmarks spanning diverse generators and scenes, together with state-of-the-art results on key explanation metrics across three benchmarks. These results are achieved with an evidence corpus roughly 10% the size of the binary corpus.

2 Related Work

Detection across image sources. Research on AI-generated image detection addresses generalization through both transferable representations and diverse training data. UniFD (Ojha et al., 2023) uses pretrained vision-language features for detection, while Effort (Yan et al., 2024) preserves pretrained knowledge while learning forgery-related patterns. Community Forensics (Park & Owens,

2025) examines the role of training diversity, showing that increasing the number and diversity of training generators improves detection on unseen generators.

Supervision for explainable detection. MLLM-based detectors learn to explain authenticity decisions from evidence annotations. FakeVLM (Wen et al., 2026) uses category-specific artifact descriptions, whereas FakeShield (Xu et al., 2025) and FakeXplainer (Ji et al., 2026) connect explanations to annotated forgery regions. Detailed explanations need not accompany the entire training corpus: SIDA (Huang et al., 2025a) adds textual supervision for a small subset after broader detection and localization training. CLOVER studies the allocation of supervision within this setting. Binary labels expand detection coverage across forgery sources, while a smaller evidence corpus emphasizes scene coverage for explanation learning.

Reinforcement learning and evidence verification. VERITAS (Tan et al., 2026) evaluates reflection quality with an image-aware reward model without reference explanations. Defake-o3 (Deng et al., 2026) and PAVE (Guo et al., 2026) assess localized evidence but also use annotated regions for reward computation. PAVE first trains evidence grounding and then adds judge-guided region–explanation alignment, while FakeXplainer (Ji et al., 2026) progressively increases the grounding reward weight. At the claim level, FaithScore (Jing et al., 2024) evaluates atomic facts against the image without reference answers. For image captioning, ClaimDiff-RL (Li et al., 2026a) constructs rewards from image-verified differences between generated and reference captions. CLOVER judges individual claims against the image and groups overlapping observations to avoid rewarding the same evidence more than once. Neither RL stage requires reference explanations or region annotations.

3 METHOD

CLOVER assigns broad source coverage to binary supervision and visual explanation learning to a smaller evidence corpus. Atomic evidence reinforcement learning then improves explanations through direct visual verification (Figure 1). Given an RGB image x, the shared MLLM can return an authenticity decision $\hat{y} \in \{\text{real}, \text{fake}\}$ and an explanation E. We denote the ground-truth label by y. An explanation with evidence takes the form $E = (e_1, \ldots, e_N)$, where each claim e_i describes a visual observation and its relevance to the decision. Different claims may describe the same observation. In our training data, the fake class includes full synthesis and generative editing; conventional non-generative edits belong to the real class.

3.1 COVERAGE-LED MULTI-TASK TRAINING

We organize detection and explanation learning into three complementary tasks that train a shared model during both SFT and RL.

Task	Input	Requested output
Binary detection	Image	Real-or-fake decision
Explainable detection	Image	Decision and visual evidence
Label-conditioned explanation	Image and authenticity label	Supplied label and visual evidence

Explainable detection connects the authenticity decision to its explanation. Supplying the label in the third task focuses training on finding and describing evidence for a known decision.

3.1.1 DATA CONSTRUCTION

Binary supervision for source coverage. To cover diverse forgery sources, we collect fully generated and generatively edited images across generators and scenes, together with real images from multiple acquisition domains. The class-balanced corpus $\mathcal{D}_b$ contains 100K images covering 35 generator groups. New sources can be added with binary labels alone. The source composition is summarized in Appendix A.5.

Evidence supervision for scene coverage. We allocate evidence annotations across scenes because the relevant cues depend on their objects and visual structures. For example, documents may show

malformed characters, while product images may show inconsistent geometry or materials. These cues can recur across generators, so we prioritize scene diversity for detailed annotation. This yields an evidence corpus $\mathcal{D}_u$ of 11K unique images spanning seven scene domains.

To create evidence annotations, two MLLMs independently describe evidence in each image without seeing its real-or-fake label. One model then reviews the evidence, removes repetition, and writes a training explanation using the known label. Finally, it checks the explanation's descriptions of visible content for unsupported claims, exaggeration, and repetition. Annotation and review details are provided in Appendix A.2.

3.1.2 JOINT SUPERVISED FINE-TUNING

We use all 11K annotated images for explainable detection and reuse a fixed 4K subset for label-conditioned explanation. Together with 100K binary examples, this yields 115K training presentations without requiring additional evidence annotations. The three tasks jointly train a shared model:

$$\mathcal{L}_{\mathrm{SFT}} = \lambda_b \mathcal{L}_b + \lambda_u \mathcal{L}_u + \lambda_c \mathcal{L}_c, \tag{1}$$

where $\mathcal{L}_b$, $\mathcal{L}_u$, and $\mathcal{L}_c$ are the standard autoregressive losses for binary detection, explainable detection, and label-conditioned explanation, respectively. The weights follow the task proportions so that each training presentation contributes equally. The loss definitions and training details are provided in Appendix A.3.

3.2 ATOMIC EVIDENCE REINFORCEMENT LEARNING

SFT teaches the model to produce authenticity decisions and explanations, but correct decisions can still accompany unsupported claims. Reference explanations may also omit valid observations. Atomic evidence RL therefore continues from SFT by assessing claims against the image and rewarding distinct supported observations. It requires only image–label pairs (x, y), without reference explanations.

3.2.1 REWARD DESIGN

Claim-level visual support. The trainable detector serves as the Actor, with policy π_θ. To assess each claim, we employ a frozen copy of the SFT model as the self-judge. For each claim e_i, the self-judge receives only (x, e_i). It assesses visual support and forensic direction: whether the observation supports AI involvement, supports no AI involvement, or has a neutral or uncertain implication.

We combine these judgments with the known label y to assign a state $q_i \in \{\mathrm{S}, \mathrm{U}, \mathrm{H}\}$: supported and label-aligned, uncertain, or hallucinated/invalid. The ground-truth label, model decision, and other claims are withheld from the self-judge. Invalid claims include visually refuted statements, claims supporting the wrong label, and claims requiring information outside the image. The full mapping, including atomicity requirements, is provided in Appendix B.2.

Correction for repeated observations. Counting claims directly can reward multiple descriptions of the same cue. A frozen Group Judge partitions the claim indices $\{1, \ldots, N\}$ into C semantic groups $\mathcal{G} = \{G_1, \ldots, G_C\}$ to avoid counting overlapping observations more than once. It receives only claim texts and identifiers. Claims connected by a shared observation belong to one group, including connections through compound statements. Grouping preserves every original claim and its state q_i.

We then aggregate the claim states with precedence $\mathrm{H} \succ \mathrm{S} \succ \mathrm{U}$: a group is invalid if any member is invalid; otherwise, it is supported if any member is supported, and uncertain if all members are uncertain. The corresponding sets of groups are

$$\begin{aligned} \mathcal{H} &= \{G_g \in \mathcal{G} : \exists i \in G_g,\ q_i = \mathrm{H}\}, \\ \mathcal{S} &= \{G_g \in \mathcal{G} \setminus \mathcal{H} : \exists i \in G_g,\ q_i = \mathrm{S}\}, \\ \mathcal{U} &= \mathcal{G} \setminus (\mathcal{S} \cup \mathcal{H}). \end{aligned} \tag{2}$$

A supported claim and an uncertain paraphrase about the same malformed character therefore contribute one supported evidence group. Figure 2 illustrates the verification and grouping process.

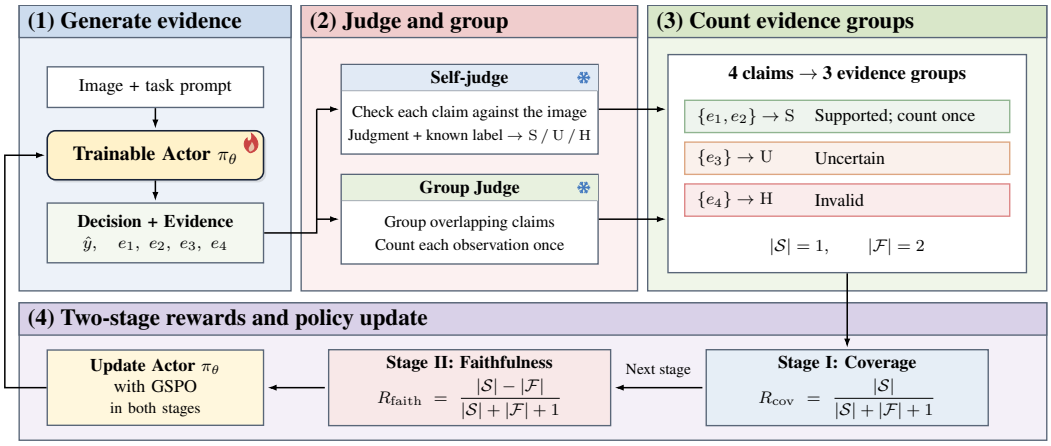


Figure 2: **Atomic evidence verification and two-stage rewards.** For a valid evidence response with a correct decision, self-judge outputs and the known label determine each claim's state. The Group Judge groups claims with overlapping observations: here, e_1 and e_2 describe the same supported cue, so four claims form three groups. The two stages sequentially update only the Actor using group sequence policy optimization (GSPO) (Zheng et al., 2025). Other response types follow Eq. (4).

Evidence rewards. Let $\mathcal{F} = \mathcal{U} \cup \mathcal{H}$ collect uncertain or invalid groups, and let $|\cdot|$ denote set cardinality. The evidence coverage and visual faithfulness rewards are

$$R_{\text{cov}}(E) = \frac{|\mathcal{S}|}{|\mathcal{S}| + |\mathcal{F}| + 1}, \qquad R_{\text{faith}}(E) = \frac{|\mathcal{S}| - |\mathcal{F}|}{|\mathcal{S}| + |\mathcal{F}| + 1}. \tag{3}$$

The constant 1 ensures that an additional supported group increases the reward even when all reported observations are supported. The marginal reward changes are derived in Appendix B.4.

Unified response reward. We combine evidence quality with decision correctness and output validity. Let $\tau \in \{b, u, c\}$ denote binary detection, explainable detection, and label-conditioned explanation. Define $v(a, \tau, y) = 1$ when response a gives the correct authenticity decision and satisfies the task's output requirements, and $v = 0$ otherwise. When visual evidence is insufficient, the model may return a single abstention notice (*Notice*) instead of evidence claims; the authenticity decision is still required.

For stage k, let $r_k = R_{\text{cov}}$ in Stage I and $r_k = R_{\text{faith}}$ in Stage II. The reward for a scored response is

$$R_k(a, \tau, y) = \begin{cases} -1, & v = 0, \\ +1, & v = 1,\ \tau = b, \\ 0, & v = 1,\ \tau \neq b,\ a \text{ is a Notice response}, \\ r_k(E), & v = 1,\ \tau \neq b,\ a \text{ contains evidence claims}. \end{cases} \tag{4}$$

Response checks and judging-failure handling are detailed in Appendices B.1 and A.4, respectively.

3.2.2 Two-stage learning

Stage I uses R_{cov} to encourage evidence discovery. Additional supported observations increase the reward, while unsupported additions already lower it when $|\mathcal{S}| > 0$. Stage II continues from this checkpoint with R_{faith}. Subtracting $|\mathcal{F}|$ from the numerator strengthens penalties for uncertain and invalid observations while retaining the incentive to report supported evidence.

For valid responses with correct decisions, explanations earn higher rewards than abstention when $|\mathcal{S}| > 0$ in Stage I and $|\mathcal{S}| > |\mathcal{F}|$ in Stage II. The second stage therefore applies a more selective criterion for rewarding explanations over abstention.

Both stages update only the Actor using GSPO (Zheng et al., 2025). GSPO compares responses to the same prompt, whereas semantic grouping combines overlapping claims within each response.

Reward normalization is analyzed in Appendix B.5. Both judges remain frozen and are used solely during training. At inference, binary detection and explainable detection require only the image and task prompt; neither judge is used.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training. We fully fine-tune Qwen3-VL-8B-Instruct (Bai et al., 2025) for one epoch on 115K training presentations (Section 3.1). This checkpoint initializes the Actor and self-judge; Qwen3-32B is the Group Judge (Yang et al., 2025a). RL uses 12,800 additional images with authenticity labels only. Each stage makes one pass over a disjoint 6,400-image set through 100 GSPO updates of 64 image–prompt pairs. Further settings are in Appendix A.

Detection benchmarks. We evaluate on GenImage (Zhu et al., 2023), LOKI-Det (Ye et al., 2025), SoFake-OOD (v3) (Huang et al., 2025b), FakeBench (Li et al., 2025), and AIGIBench (Li et al., 2026b). We average their macro-F1 scores equally; GenImage and AIGIBench first average over generator or source groups. Separate tests report macro-F1 and balanced accuracy on Nano-banana-150K (BitMind, 2025), and fake recall on the fake-only GPT-Image2-wild set (Zewde et al., 2026).

Explanation benchmarks. TriDF (Jiang-Lin et al., 2026) compares evidence with reference artifact sets. Type A supplies known-fake labels; Type B requests detection and explanation on mixed real/fake images without labels. Both score evidence only on fake images using precision-weighted $F_{0.5}$. LOKI-Explain (Ye et al., 2025) supplies known-fake labels and reports an Overall score (0–100). On FakeClue (Wen et al., 2026), our FakeClue-Evidence protocol checks claims against fake images, including valid observations absent from references. Precision and Hallucination Rate average the per-image fractions of supported and visually hallucinated semantic groups. Evi@Img is the fraction of images with supported evidence; Sup/Img is the mean number of supported groups per image.

Comparison protocol. We compare general-purpose MLLMs (Wang et al., 2025; Bai et al., 2025), binary detectors (Ojha et al., 2023; Cheng et al., 2024; 2025; Yang et al., 2025b; Yan et al., 2024), and explanation-capable detectors (Xu et al., 2025; Huang et al., 2025a; Wen et al., 2026; Xia et al., 2025; Tan et al., 2026) (Table 1) using fixed samples, parsers, and scorers. CLOVER and general-purpose MLLMs share a detection prompt; specialized detectors use their released interfaces. A fixed Qwen3.8-Max evaluator, separate from the RL judges, scores explanations. Some generator families overlap with training; AIGIBench uses a disjoint split of a training source corpus. Appendix C details these relationships and our protocols, which differ from the original leaderboards. Unit-interval evaluation metrics are multiplied by 100; other quantities retain their original scales.

4.2 MAIN RESULTS

Leading average detection performance. CLOVER-SFT and the final RL model reach 92.44% and 92.35% average macro-F1, exceeding the strongest competitor, FakeVLM, by 6.13 and 6.04 percentage points (Table 1). Gains are largest on SoFake-OOD and AIGIBench; FakeVLM remains ahead on GenImage and LOKI-Det. The final model also leads on Nano-banana-150K with 80.12% macro-F1 and 84.22% balanced accuracy, but its 85.33% GPT-Image2-wild fake recall trails ProDet's 90.53% (Table 9).

RL improves AIGIBench but lowers the mean by 0.09 points through small declines elsewhere. SFT thus establishes the detection advantage with limited evidence annotations, while RL largely preserves it as explanations improve.

More reliable explanations without new reference annotations. RL improves explanation quality across all three benchmarks without additional reference explanations (Table 2). FakeClue-Evidence Precision rises from 54.84 to 61.46, while Hallucination Rate falls from 14.55 to 9.29. Higher Evi@Img and Sup/Img show that precision improves alongside evidence coverage. TriDF $F_{0.5}$ improves under both prompt settings; LOKI-Explain Overall rises from 67.69 to 73.52.

Table 1: **Detection macro-F1 (%) on five benchmarks.** Avg.: unweighted mean. Across result tables, bold/underlining mark best/second-best; green/red mark improvement (or ties)/decline. Yellow highlights Stage II; gray marks other proposed configurations.

Method	GenImage	LOKI-Det	SoFake-OOD	FakeBench	AIGIBench	Avg.↑
General-purpose MLLMs						
InternVL3.5-8B	64.13	56.11	49.42	77.14	63.90	62.14
InternVL3.5-38B	58.33	51.20	55.42	73.91	65.12	60.80
Qwen3-VL-8B-Instruct	72.83	65.64	78.40	84.43	73.10	74.88
Qwen3-VL-32B-Instruct	62.62	57.33	73.04	76.51	70.34	67.97
Binary detectors						
UniFD (CVPR'23)	57.90	43.91	33.74	63.92	51.14	50.12
ProDet (NeurIPS'24)	36.53	40.22	38.92	35.61	45.13	39.28
CO-SPY (CVPR'25)	75.84	59.60	63.11	85.42	70.73	70.94
D^3 (CVPR'25)	97.91	74.22	72.72	86.01	78.23	81.82
Effort (ICML'25)	76.11	61.53	53.80	71.43	57.63	64.10
Explanation-capable detectors						
FakeShield (ICLR'25)	87.33	58.60	80.61	88.71	62.52	75.55
SIDA (CVPR'25)	54.01	46.02	61.41	62.21	56.31	55.99
FakeVLM (NeurIPS'25)	**99.21**	**84.60**	74.51	94.20	79.02	86.31
MIRAGE-R1 (AAAI'26)	59.44	52.80	61.72	62.70	52.81	57.89
VERITAS (ICLR'26)	55.91	40.31	38.43	57.73	45.31	47.54
Ours						
CLOVER-SFT	<u>98.11</u>	<u>83.94</u>	**93.72**	**96.72**	89.71	**92.44**
CLOVER-RL (Stage I)	98.01	83.30	<u>93.70</u>	<u>96.45</u>	<u>90.62</u>	<u>92.42</u>
CLOVER-RL (Stage II)	97.93	83.00	<u>93.70</u>	96.10	**91.01**	92.35

Stage I raises FakeClue-Evidence Evi@Img from 65.48 to 70.43; Stage II maintains this coverage while further improving precision and reducing hallucinations. TriDF $F_{0.5}$ improves at both stages, although LOKI-Explain Completeness decreases slightly in Stage II. Full metrics are in Appendices D.2 and D.3.

The final model leads the evaluated methods on TriDF $F_{0.5}$ and FakeClue-Evidence Precision. Qwen3-VL-32B remains stronger on LOKI-Explain Overall, and FakeVLM provides supported evidence for more FakeClue images. Blinded ratings on 300 fake images also find higher precision and lower hallucination rates after RL (Appendix E.1).

4.3 Ablation studies

Complementary supervision. At matched SFT update counts, binary-only training yields 93.12% detection macro-F1 and 31.76 FakeClue-Evidence Precision; evidence-only training yields 71.18% and 49.33, respectively. Joint training retains 92.44% detection macro-F1 and improves precision to 54.84 with an evidence corpus roughly 10% the size of the binary corpus (Table 3(a)). Binary supervision thus sustains detection, while evidence annotations improve explanations.

Binary source coverage. At fixed corpus sizes, restricting generators, fake scenes, or real-image acquisition domains in the binary corpus lowers detection but barely changes evidence precision (Table 3(b)). Detection therefore benefits from source diversity beyond data quantity. With evidence annotations fixed, binary adaptation to new sources raises their macro-F1 from 76.80% to 85.30%, versus 75.80% with matched updates on old sources (Appendix E.2). This supports source expansion without new evidence annotations.

Scene coverage. At a fixed evidence budget, expanding from two to seven scenes raises FakeClue-Evidence Precision from 47.64 to 54.84 with little change in detection (Table 3(c)). Expanding from

Table 2: **Explanation quality (0–100).** Differences are absolute; a lower Hallucination Rate is better. LOKI denotes LOKI-Explain. Formatting follows Table 1.

Method	TriDF $F_{0.5}$ ↑		LOKI	FakeClue-Evidence		
	Type A	Type B	Overall↑	Evi@Img↑	Precision↑	Hall. Rate↓
General-purpose MLLMs						
InternVL3.5-8B	39.33	13.60	55.77	26.03	14.42	12.18
InternVL3.5-38B	32.39	7.08	53.73	43.36	29.92	9.39
Qwen3-VL-8B-Instruct	34.63	20.44	70.47	28.82	22.22	11.74
Qwen3-VL-32B-Instruct	44.77	27.40	**76.78**	32.61	25.52	12.08
Explanation-capable detectors						
FakeShield (ICLR'25)	11.67	1.76	36.70	3.45	1.80	12.41
FakeVLM (NeurIPS'25)	22.33	18.47	59.06	**74.87**	34.33	25.45
MIRAGE-R1 (AAAI'26)	46.67	26.69	55.41	13.19	7.77	15.10
VERITAS (ICLR'26)	43.89	25.87	52.90	47.53	27.79	12.97
Ours						
CLOVER-SFT	43.47	36.15	67.69	65.48	54.84	14.55
CLOVER-RL (Stage I)	45.80	38.65	70.39	70.43	59.05	13.42
CLOVER-RL (Stage II)	**48.62**	**41.01**	73.52	70.43	**61.46**	**9.29**
Δ (Stage II − SFT)	**+5.15**	**+4.86**	**+5.83**	**+4.95**	**+6.62**	**−5.26**

Table 3: **Core ablations** with matched updates (a), fixed data budgets (b–c), or different claim judges (d). Det.: five-benchmark mean macro-F1 (%). FC Prec.: FakeClue-Evidence Precision (0–100).

(a) Supervision combination

Supervision	Det.↑	FC Prec.↑
Binary only	**93.12**	31.76
Evidence only	71.18	49.33
Binary + Evidence	92.44	**54.84**

(c) Evidence coverage

Scenes / generators	Det.↑	FC Prec.↑
2 / 4	**92.44**	47.64
7 / 4	**92.44**	54.84
7 / 20	92.41	**54.98**

(b) Binary source coverage

Allocation	Det.↑	FC Prec.↑
Narrow generators	84.68	53.56
Narrow fake scenes	85.23	53.18
Natural-real only	85.57	54.59
Full allocation	**92.44**	**54.84**

(d) Choice of claim judge

Judge	Det.↑	FC Prec.↑
Dynamic self-judge	92.32	56.16
Qwen3-VL-8B	92.33	57.17
Qwen3-VL-32B	92.28	58.09
Self-judge (ours)	**92.35**	**61.46**

four to twenty generators within seven scenes adds little. Scene diversity matters more under this budget, consistent with visual cues shared across generators.

Choice of claim judge. At similar detection macro-F1, the fixed self-judge yields 61.46 FakeClue-Evidence Precision after RL, versus 58.09 with Qwen3-VL-32B (Table 3(d)). Refreshing the self-judge from the Actor lowers precision. Together with strong SFT detection (Table 1), these results suggest that task specialization can provide useful feedback in this setup.

Effect of reward order. At matched update counts, coverage followed by faithfulness leads the tested schedules on TriDF-B $F_{0.5}$ and FakeClue-Evidence Precision (Appendix D.5), supporting evidence discovery before stricter evidence selection.

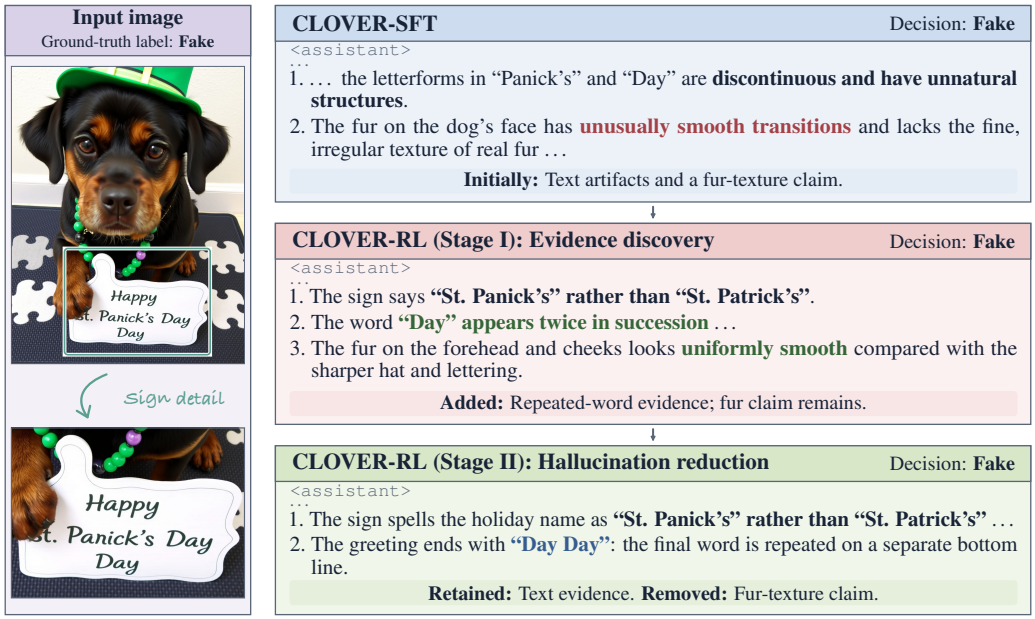


Figure 3: **Evidence discovery and refinement.** Stage I identifies more specific text artifacts, including misspelling and word repetition. Stage II retains this evidence and removes the less certain fur-texture claim, while the authenticity decision remains unchanged.

Qualitative analysis. On the same image, SFT cites both text artifacts and smooth fur (Figure 3). Stage I identifies the misspelled holiday name and repeated "Day". Stage II retains these text observations and drops the fur claim. The decision remains fake throughout, while evidence becomes more specific and selective. Additional examples, including a Notice response, are shown in Appendix E.4.

5 Limitations

CLOVER uses automated evidence annotations and judgments. Its self-judge may share detector errors because both originate from the SFT model. Visual verification requires identifiable image cues, which some synthetic images may lack. The model may abstain despite available evidence (Appendix E.3). Explanation evaluation is primarily automated; independent blinded assessment covers 300 images (Appendix E.1). Some evaluation sets share training generator families or source corpora (Table 7), so they do not test entirely unseen sources. Small differences between controls need confirmation across independent runs. The evidence budget measures corpus size; annotation time and cost were not measured.

6 Conclusion

We presented CLOVER, a training framework that jointly optimizes three complementary tasks to combine broad binary supervision with a smaller, scene-diverse evidence corpus. The framework extends detection coverage to new forgery sources without requiring additional explanation annotations. Atomic evidence reinforcement learning uses a self-judge to assess claims against the image and rewards distinct supported observations. Its two stages encourage evidence discovery and then strengthen penalties for unsupported or uncertain claims. With an evidence corpus roughly 10% the size of the binary corpus, CLOVER achieves the best average macro-F1 across five detection benchmarks. Further evidence optimization improves explanation quality across three benchmarks and reduces hallucinations while largely preserving detection performance. Together, these results show that detection coverage and explanation reliability can improve without a corresponding expansion of detailed evidence annotations.

AI USE STATEMENT

Large language models were used for grammar and language refinement. Multimodal large language models were used for training-data annotation.

ETHICS STATEMENT

CLOVER is intended to assist human assessment of image authenticity. Incorrect judgments can misrepresent genuine content and harm its creators. Its decisions and explanations should therefore be considered alongside the image and other available evidence. We report hallucinations and distinguish evidence abstention from the authenticity decision to make the limits of automated assessment explicit.

REPRODUCIBILITY STATEMENT

Appendices A and B provide data construction, training settings, judging rules, and reward analysis. Full prompts are included in the accompanying `prompts/` directory. Evaluation protocols, complete results, and supplementary assessment procedures are provided in Appendices C, D, and E.

REFERENCES

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

BitMind. Nano-banana-150k dataset. Hugging Face Datasets, 2025. URL `https://huggingface.co/datasets/bitmind/Nano-banana-150k`. Accessed: 2026-09-26.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37:21979–21998, 2024.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13455–13465. IEEE, 2025.

Bowen Deng, Jiahui Zhan, Yikun Ji, Haozhen Yan, and Jianfu Zhang. Defake-o3: From speculative rationales to verifiable evidence for explainable aigi detection. *arXiv preprint arXiv:2608.16259*, 2026.

Kun Guo, Yuzhou Yang, Haoyue Wang, Qichao Ying, Sheng Li, and Zhenxing Qian. Grounding and explaining visual evidence for ai-generated image detection in human-centric scenes. *arXiv preprint arXiv:2608.01988*, 2026.

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. IEEE, 2025a.

Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, et al. So-fake: Benchmarking and explaining social media image forgery detection. *arXiv preprint arXiv:2505.18660*, 2025b.

Yikun Ji, Yan Hong, Qi Fan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, et al. Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 68549–68570, 2026.

Jian-Yu Jiang-Lin, Kang-Yang Huang, Ling Zou, Ling Lo, Sheng-Ping Yang, Yu-Wen Tseng, Kun-Hsiang Lin, Chia-Ling Chen, Yu-Ting Ta, Yan-Tsung Wang, et al. Tridf: Evaluating perception, detection, and hallucination for interpretable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17087–17098, 2026.

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, 2024.

Tianle Li, Xuyang Shen, Yan Ma, Rongxin Guo, Shaoxiang Chen, Jiacheng Chen, Haochen Wang, Hongyang Tang, Yucong Zhou, and Yu Cheng. Claimdiff-rl: Fine-grained caption reinforcement learning through visual claim comparison. *arXiv preprint arXiv:2605.20278*, 2026a.

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

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *Advances in neural information processing systems*, 38, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257. IEEE, 2025.

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, 2018.

Hao Tan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, Zhen Lei, et al. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 66420–66477, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026.

Cheng Xia, 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. *arXiv preprint arXiv:2508.13223*, 2025.

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.

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. *arXiv preprint arXiv:2411.15633*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D 3: Scaling up deepfake detection by learning from discrepancy. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23850–23859. IEEE, 2025b.

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

Kidus Zewde, Simiao Ren, Xingyu Shen, Jiaqi Wu, Yuchen Zhou, Tommy Duong, Zikang Zhang, Ethan Traister, and Kewen Xie. Gpt-image-2 in the wild: A twitter dataset of self-reported ai-generated images from the first week of deployment. *arXiv preprint arXiv:2604.25370*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

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. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Data and Training Details

A.1 Training data

The binary corpus contains 50K real and 50K fake images, covering 35 generator groups, full synthesis and generative editing, seven fake-scene domains, and three real-image acquisition domains.

The evidence corpus contains 11K unique images: a 10K core and 1K additional images with conventional, non-generative edits. These additional images are labeled real under our definition of AI involvement. All 11K images are used for explainable detection, and a fixed 4K subset is also used for label-conditioned explanation. This produces 15K evidence presentations and 115K presentations overall. A presentation consists of an image paired with a task prompt; reusing an image for another task does not require another evidence annotation.

The seven scene domains in the evidence corpus cover humans and actions; natural scenes; products and materials; text, documents, and interfaces; illustration and design; aerial and satellite images; and medical and scientific content. Source, generation mode, scene, and acquisition-domain labels are recorded to construct the coverage comparisons. Related frames, edits, and crops share a content-group identifier and remain within the same training or evaluation split. Table 7 specifies the relationship between training and each evaluation set, including any overlap in generator families.

A.2 Evidence annotation and review

We fix authenticity labels and source identifiers before annotation. GPT-5.5 and Claude Opus 4.8 independently inspect each RGB image without its label to obtain complementary candidate observations. Claude Opus 4.8 then combines these observations with the known label to produce a training explanation.

The annotations distinguish two types of evidence. *Visible-semantic evidence* (VSE) describes an inspectable object, attribute, or relation. *Qualified low-level evidence* (QLE) describes an image-specific low-level hypothesis using appropriately tentative language. A separate Claude Opus 4.8 review checks VSE for visual support, exaggerated wording, and repetition without introducing new observations. QLE does not undergo this visual truth check because such hypotheses are difficult to verify reliably from the image alone. They must remain qualified and must not imply access to unavailable measurements or hidden model activations. This distinction applies to annotation review; generated claims are subsequently assessed during RL using the self-judge rules in Appendix B.2.

Both evidence subsets undergo automatic checks for label consistency, output format, image–source correspondence, consecutive numbering, and the requirement that a Notice be the sole numbered item. Training targets contain the authenticity decision followed by the retained evidence or a Notice. Source metadata and annotation-review records are excluded from the targets. The full annotation instructions are provided in `prompts/annotation/`.

A.3 Supervised training

We fine-tune all parameters of Qwen3-VL-8B-Instruct for one epoch over the 115K presentations using a 4,096-token context. AdamW (Loshchilov & Hutter, 2017) uses a learning rate of 5×10^{-6}, weight decay of 0.1, cosine decay, and 1% warm-up. Training uses bfloat16 on 16 data-parallel GPUs, with a micro-batch of one and gradient accumulation over 16 steps, giving a global batch size of 256.

For image x, task prompt t, and target response a, the per-presentation loss is

$$\ell_\theta(x, t, a) = -\frac{1}{|a|} \sum_{j=1}^{|a|} \log \pi_\theta(a_j \mid x, t, a_{<j}). \tag{5}$$

Here a_j is the jth assistant target token and $|a|$ is the target length. System, user, image-placeholder, and padding tokens are masked.

Let $\mathcal{D}_c \subset \mathcal{D}_u$ denote the fixed 4K subset reused for label-conditioned explanation. The prompts t_b, t_u, and $t_c(y)$ specify the three tasks, with $t_c(y)$ supplying the authenticity label. The task losses are

$$\begin{aligned} \mathcal{L}_b &= \mathbb{E}_{(x,y)\sim\mathcal{D}_b}[\ell_\theta(x, t_b, y)], \\ \mathcal{L}_u &= \mathbb{E}_{(x,y,E)\sim\mathcal{D}_u}[\ell_\theta(x, t_u, (y, E))], \\ \mathcal{L}_c &= \mathbb{E}_{(x,y,E)\sim\mathcal{D}_c}[\ell_\theta(x, t_c(y), (y, E))]. \end{aligned} \tag{6}$$

Here (y, E) denotes the serialized response containing the decision and explanation, including the repeated label in label-conditioned explanation. Labels in the task losses are serialized as the yes/no answers required by the prompt's language and question polarity. Each task loss averages the per-presentation loss uniformly over its examples. The weights in Eq. (1) follow the presentation counts:

$$\lambda_b = \frac{100}{115}, \qquad \lambda_u = \frac{11}{115}, \qquad \lambda_c = \frac{4}{115}. \tag{7}$$

The joint objective therefore gives the mean loss over all 115K presentations; these weights are determined by the data proportions rather than tuned separately.

Conditional factorization and task roles. Given an image x and an explainable-detection prompt t, the model generates an authenticity decision $\hat{y}$ followed by an explanation E, which can be a Notice when evidence is insufficient. The output distribution follows the chain rule:

$$\pi_\theta(\hat{y}, E \mid x, t) = \pi_\theta(\hat{y} \mid x, t)\, \pi_\theta(E \mid x, \hat{y}, t). \tag{8}$$

The two factors describe deciding whether an image is real or fake and explaining that decision. This identity holds for a given prompt; it does not equate the conditional distributions under the three task prompts.

When computing explanation-token losses, both explanation tasks receive the ground-truth decision prefix y through teacher forcing. Label-conditioned explanation also provides the label in the input prompt before generation. Its target repeats the supplied label, and these output tokens contribute to the loss along with the explanation tokens.

Supervision and coverage comparisons. The single-supervision controls replay the remaining data to match the primary SFT update count. To test label-conditioned explanation, both variants reuse the same 4K images and differ only in whether the additional prompt supplies the authenticity label.

Evidence-coverage comparisons use 11K annotated images and 15K presentations. Scene-coverage variants match class balance, generator groups, evidence types, language, and training updates. The generator-coverage comparison retains all seven scenes while varying the generator groups represented in the evidence corpus.

Binary-coverage comparisons keep 100K images and preserve class balance and the proportion of fully generated versus generatively edited images. Each comparison restricts one aspect of the binary corpus: generator groups, fake scenes, or real-image acquisition domains. The evidence corpus remains unchanged. Let (G_B, D_F, D_R) denote the numbers of binary generator groups, fake-scene domains, and real acquisition domains. The narrow-generator, narrow-fake-scene, natural-real-only, and full configurations use $(8, 7, 3)$, $(35, 2, 3)$, $(35, 7, 1)$, and $(35, 7, 3)$, respectively. The evidence corpus's four generator groups are included in the narrow eight binary groups, which are themselves included in the full 35 groups. The shared evidence corpus can still contain scenes or acquisition domains absent from a restricted binary corpus.

Binary corpus size. The scale comparison uses nested pools,

$$\mathcal{D}_b^{50K} \subset \mathcal{D}_b^{100K} \subset \mathcal{D}_b^{200K},$$

with matched source, generator, scene, and class proportions. Each pool is combined with the same 15K evidence presentations from 11K annotated images. One epoch produces 254, 450, and 840 optimizer updates, respectively. This comparison therefore measures the combined effect of binary data volume and the associated one-epoch training budget.

Extra SFT continuation. This control tests whether additional supervised training can reproduce the gains from RL. It starts from the primary SFT checkpoint and adds 200 supervised updates. Presentations are sampled uniformly with replacement from the original 100K binary detection, 11K explainable detection, and 4K label-conditioned explanation examples, without per-batch task quotas.

The targets, global batch size of 256, context length, numerical precision, and loss remain unchanged, resulting in 51,200 replay visits. AdamW moments and the learning-rate schedule restart, with two warm-up updates to 5×10^{-6} followed by 198 updates of cosine decay to zero. Weight decay remains 0.1. We evaluate the checkpoint after 200 updates. This matches the nominal update count of RL, while the numbers of examples, generated tokens, and judging calls differ.

A.4 Reinforcement learning

The primary SFT checkpoint initializes the Actor and self-judge. RL uses 12,800 additional training images that are not part of the SFT image sets. Only their authenticity labels are used; reference explanations are not required. The images are divided into two disjoint stage-specific sets of 6,400 images. Each stage visits every image–prompt pair once, giving one epoch and 100 GSPO updates. An update contains 64 prompts with eight sampled responses per prompt, giving 512 accepted responses. Decoding uses temperature 0.9, top-p 0.95, top-k 50, and a maximum response length of 1,024 tokens. The Actor uses a learning rate of 10^{-7}, weight decay of 0.01, five warm-up updates, and a Kullback–Leibler (KL) divergence penalty with coefficient 0.03. The KL term enters the policy loss separately from reward standardization.

SFT and RL use the same three task prompts. Each RL update contains 16 binary detection, 32 explainable detection, and 16 label-conditioned explanation prompts. Each task balances language and authenticity class. Explainable detection also balances direct and reverse question polarity, while label-conditioned explanation balances its two prompt variants.

Qwen3-32B serves as the frozen Group Judge. If a required self-judge output is missing or the Group Judge returns an invalid partition, all eight responses for that prompt are replaced. Reward standardization is defined in Appendix B.5.

Fixed and dynamic self-judges. The default self-judge retains the SFT checkpoint throughout RL. The dynamic self-judge instead copies the current Actor before response sampling at each accepted optimizer update. This snapshot remains fixed throughout that update, including any replacement responses. Judging inputs, prompts, decoding, state mapping, the Group Judge, and the training budget match the fixed self-judge condition. Gradients update only the Actor. This control tests the effect of allowing the judge to change with the Actor.

Reward schedules. All five schedules start from the primary SFT checkpoint and use the same ordered image–prompt pairs, including two disjoint blocks of 100 updates. Single-objective schedules use one reward for all 200 updates. Ordered schedules use each reward for 100 updates. The mixed schedule uses

$$R_{\text{mix}} = \frac{R_{\text{cov}} + R_{\text{faith}}}{2}$$

throughout training. Decision checks, semantic grouping, and task proportions remain unchanged. The 100-update comparisons use the first block. These controls match accepted updates; generated tokens and judging calls can vary across schedules. Reward-component and task-mixture comparisons are described alongside their results in Appendix D.5.

A.5 Generator groups

The 35 generator groups used for binary supervision are shown in Table 4. Related model versions are combined where indicated, and each group serves as one unit in the source-coverage comparisons.

Table 4: Generator groups in the binary training corpus. Related model versions are combined into 35 groups as listed.

Group IDs	Count	Generator groups
1–5	5	DALL-E 2; Midjourney V1–V4; Stable Diffusion 1.x; Imagen; NovelAI Diffusion
6–11	6	Midjourney V5/5.1/5.2; DALL-E 3; Adobe Firefly; SDXL 1.0; Ideogram 0.1; DeepFloyd IF
12–19	8	FLUX.1; Midjourney V6/V6.1; SD3/3.5; Imagen 3; Ideogram 1/2; Firefly Image 3; Recraft V2/V3; Kolors
20–27	8	GPT Image 1/1.5; Nano Banana/Pro; Midjourney V7; Imagen 4; Seedream 4/4.5; FLUX.1 Kontext / FLUX.2; Qwen-Image; Ideogram 3
28–35	8	GPT Image 2; Nano Banana 2; Midjourney V8/8.1; Seedream 5.0 Lite/Pro; Muse Image; Ideogram 4.0; Recraft V4/V4.1; FLUX.2 [klein]

B Evidence Judgments and Reward Analysis

B.1 Task prompts and response checks

The instructions in `prompts/task_instructions/` implement the three tasks in Section 3.1 and are shared by SFT and RL. The following rules specify how responses are parsed and checked before judging.

Decision and explanation format. The explainable detection prompt asks whether generative AI generated or edited any part of the image. Conventional non-generative editing does not count as AI involvement. The response begins with yes or no, followed by `Reasons:` and consecutively numbered claims. Each claim should describe a visual observation and its relevance to the authenticity decision.

A reverse-polarity prompt asks whether generative AI played no part in the image. The parser accounts for this reversal when converting yes or no into an authenticity decision. English and Chinese prompts follow the same numbering and polarity rules, without prescribing a fixed number of claims. Label-conditioned explanation uses the same response format but requires the model to repeat the supplied label before explaining it.

Evidence abstention. The Actor may return a Notice when it cannot provide defensible visual evidence. The authenticity decision is still required. Each language and label direction has a designated Notice string, which must appear as the sole numbered item. The four strings are provided in `prompts/notice_strings.txt`. After trimming line boundaries and the numbered-item prefix, the parser requires an exact match, including case and punctuation.

Response validation. Table 5 specifies the parser checks used to determine response validity in Eq. (4). Only valid explanations containing evidence claims proceed to the self-judge and Group Judge.

B.2 Self-judge outputs and claim states

The self-judge assesses each original claim together with its RGB image. It does not receive the authenticity label, the Actor's decision, other claims, source information, semantic groups, or rewards. Its output separates visual support from the authenticity implication of the observation:

- `visual_semantic` assesses claims about objects, relations, geometry, lighting, materials, text, or visible editing.
- `pixel_forensic` assesses localized low-level phenomena, such as repeated micro-texture, smoothing, noise inconsistency, or discontinuities in image detail.

Table 5: Response checks for the three tasks. Evidence responses proceed to judging only after the authenticity and format checks pass; a valid Notice bypasses both judges.

Route and condition	**Interpretation**	**Result**
Binary: correct parsed decision	Correct decision	$+1$
Binary: wrong or unparseable decision	Invalid decision	-1
Evidence: wrong authenticity decision, conditioned-label mismatch, or forbidden control text	Invalid response	-1
Evidence: malformed header, numbering, or empty claim	Invalid format	-1
Evidence: Notice in a disallowed language, direction, position, or mixture	Invalid Notice	-1
Evidence: sole exact direction-matched Notice	Evidence abstention	0
Evidence: all gates pass with evidence claims	Evidence requires verification	Continue

- `forensic_direction` identifies whether the observation supports AI involvement, supports no AI involvement, or has a neutral or uncertain implication.

Each support field takes one of five values: `not_claimed`, `supported`, `refuted`, `uncertain`, or `external_dependency`. The direction field takes `ai_involved`, `no_ai_involved`, `neutral`, or `uncertain`.

We combine these judgments with the known authenticity label to assign $q_i \in \{\mathrm{S}, \mathrm{U}, \mathrm{H}\}$. The mapping rules in Table 6 are applied in order. Refuted claims and claims requiring unavailable information receive H, as do claims with neutral or incorrect forensic direction. The remaining rules distinguish supported atomic claims from uncertain claims and claims that combine both evidence axes. The label is used only in this mapping and is never supplied to the self-judge.

Table 6: Mapping self-judge outputs and the known label to claim states. Rules are applied from top to bottom.

Evidence-axis condition	**Direction condition**	**State**	**Interpretation**
Either axis is external-dependent or refuted	Any	H	Unavailable or image-contradicted claim
Any remaining axes	Opposite of target or neutral	H	Wrong or non-forensic direction
Support on both axes	Target-aligned or uncertain	H	Non-atomic dual-axis claim
Claims on both axes, at least one uncertain	Target-aligned or uncertain	U	In-scope but unresolved atomicity
Neither axis claims evidence	Any remaining direction	H	No concrete verifiable phenomenon
One supported axis; other unclaimed	Target-aligned	S	Atomic, supported, aligned evidence
One supported axis; other unclaimed	Uncertain	U	Supported phenomenon, unresolved direction
One uncertain axis; other unclaimed	Target-aligned or uncertain	U	Concrete but visually unresolved evidence

B.3 Semantic grouping

The Group Judge determines which claims describe overlapping observations. It receives only the original claim identifiers and texts and does not assess truth or visual support. Claim states remain those assigned by the self-judge and label mapping.

Claims are merged when they refer to the same concrete observation, including paraphrases and overlapping broad or specific descriptions. Sharing an object or anomaly type is not sufficient to merge independently assessable facts. Enumerating several instances of one repeated pattern still describes one observation.

Compound statements connect groups through shared observations. For example, statements describing A, B, and $A + B$ belong to one group because the compound statement overlaps with both. A compound statement without overlap forms its own group. The Group Judge does not split or remove claims, and compound structure alone does not change a claim's state.

The output contains `clusters`, each with `member_ids`, and must include every input identifier exactly once. Full instructions are provided in `prompts/group_judge/`.

The resulting groups are assigned to $\mathcal{S}, \mathcal{U}$, and $\mathcal{H}$ using Eq. (2).

For N claims partitioned into C groups, $N - C$ counts claims merged through repetition or overlap. Because a compound statement can connect several observations, C is the number of reward groups and need not equal the number of individual visual facts.

B.4 Marginal effects of the evidence rewards

The two rewards share an incentive to report additional supported evidence but differ in their treatment of uncertain and invalid evidence. Coverage here refers to supported evidence in the response, not recall against an exhaustive inventory of image cues. Let $T = |\mathcal{S}| + |\mathcal{F}| + 1$, and hold existing group assignments and states fixed. Adding one supported group changes the rewards by

$$\Delta_{\mathcal{S}} R_{\text{cov}} = \frac{|\mathcal{F}| + 1}{T(T+1)}, \qquad \Delta_{\mathcal{S}} R_{\text{faith}} = \frac{2|\mathcal{F}| + 1}{T(T+1)}. \tag{9}$$

Both changes are positive, including when every existing group is supported. Adding one uncertain or invalid group instead gives

$$\Delta_{\mathcal{F}} R_{\text{cov}} = -\frac{|\mathcal{S}|}{T(T+1)}, \qquad \Delta_{\mathcal{F}} R_{\text{faith}} = -\frac{2|\mathcal{S}| + 1}{T(T+1)}. \tag{10}$$

The faithfulness reward therefore applies a stronger penalty to such additions. It also penalizes them when $|\mathcal{S}| = 0$, whereas the coverage reward remains zero in that case.

Repeating an observation does not change either reward if the new claim remains in an existing group and preserves its state. An invalid paraphrase can change the group state, so this invariance applies to repeated meaning with unchanged judgments.

These properties describe raw rewards. Their effect on GSPO advantages also depends on the other responses sampled for the same prompt.

B.5 Effect of GSPO reward normalization

GSPO standardizes rewards within the responses sampled for each prompt (Zheng et al., 2025). For a reward vector $\mathbf{r}$, the advantage vector is

$$A_\delta(\mathbf{r}) = \frac{\mathbf{r} - \bar{r}\mathbf{1}}{\sigma(\mathbf{r}) + \delta}, \qquad \delta = 10^{-8}, \tag{11}$$

where $\bar{r}$ is the mean reward, $\sigma(\mathbf{r})$ is the population standard deviation, and $\mathbf{1}$ is the all-ones vector. Constant-reward groups receive zero advantage.

Consider valid evidence responses that all contain the same number $C = |\mathcal{S}| + |\mathcal{F}|$ of semantic groups. Let $\mathbf{r}_{\text{cov}}$ and $\mathbf{r}_{\text{faith}}$ collect their rewards under the two objectives. Then

$$\mathbf{r}_{\text{faith}} = 2\mathbf{r}_{\text{cov}} - \frac{C}{C+1}\mathbf{1}. \tag{12}$$

For nonconstant reward vectors,

$$A_\delta(\mathbf{r}_{\text{faith}}) = \frac{2(\sigma(\mathbf{r}_{\text{cov}}) + \delta)}{2\sigma(\mathbf{r}_{\text{cov}}) + \delta} A_\delta(\mathbf{r}_{\text{cov}}). \tag{13}$$

Adding a constant does not affect centered rewards, and multiplication by two doubles their standard deviation. Consequently, the two advantage vectors are identical when $\delta = 0$. With the implemented stabilizer, their direction remains the same and their scale differs by the factor above. A stronger penalty in the raw reward therefore need not change relative advantages within this setting.

When response j contains C_j groups, the corresponding rewards instead satisfy

$$r_{\mathrm{faith},j} = 2r_{\mathrm{cov},j} - \frac{C_j}{C_j + 1}. \tag{14}$$

The subtracted term now varies across responses and can change relative advantages. Notice and failed responses also retain fixed rewards of zero and -1 across stages, so response groups containing them need not satisfy the same affine relation. Binary detection rewards remain unchanged and are normalized within their own prompt groups.

Table 17 evaluates the resulting training schedules, complementing this analysis of fixed responses.

C Evaluation Protocols

C.1 Evaluation sets and model comparisons

Table 7 gives the size of each evaluation set and its relationship to training. We use fixed evaluation samples and scoring procedures across models. Before aggregation, we check sample identifiers, the alignment of responses with ground-truth labels, and parsing.

The comparison includes InternVL3.5 models with 8B and 38B parameters (Wang et al., 2025), Qwen3-VL-Instruct models with 8B and 32B parameters (Bai et al., 2025), five binary detectors, and five explanation-capable detectors. The Instruct suffix is omitted from Qwen3-VL names in prose and ablation tables. Detection follows the comparison protocol in Section 4.1; the benchmark-specific explanation inputs are defined below.

Table 7: Evaluation sets and their relationship to training. Content-disjoint means no shared benchmark image, prompt record, annotation, or content group. Generator families can overlap even when the content does not.

Evaluation set	**Images**	**Relationship to training**
GenImage	20,000	Content-disjoint; generator families overlap.
LOKI-Det	2,217	Content-disjoint; no LOKI source or annotations.
SoFake-OOD	20,000	Content-disjoint; no So-Fake source or annotations.
FakeBench	6,000	Content-disjoint; no FakeBench training data.
AIGIBench	20,000	Disjoint split of the same corpus represented in training.
Nano-banana-150K	20,000	Content-disjoint; independently collected family data used.
GPT-Image2-wild	10,217	Content-disjoint; independently collected family data used.
TriDF A/B	1,402/2,542	Content-disjoint; no TriDF source or annotations. The two suites share their fake subset.
LOKI-Explain	229	Content-disjoint; no LOKI source or annotations.
FakeClue-Evidence	5,000	No evaluation images used; related FakeVLM training data used.

SIDA (Huang et al., 2025a) is included in detection comparisons. We omit its explanation results because our reproduction of its multi-module explanation pipeline repeatedly failed the required instruction-following and output-format checks.

Before computing the five-benchmark detection average, we average GenImage scores across its eight generators and AIGIBench scores across its 25 sources. SoFake-OOD treats both tampered and fully synthetic images as fake. Table 7 also gives the sample sizes for the two additional detection tests defined in Section 4.1.

Explanation scoring uses a fixed Qwen3.8-Max evaluator through Alibaba Cloud Model Studio, with temperature zero, thinking disabled, and structured JSON output. Benchmark-specific prompts,

parsers, and aggregation rules remain fixed across evaluated models. The corresponding instructions are provided in `prompts/evaluation/`.

C.2 TRIDF ARTIFACT-SET EVALUATION

TriDF Type A supplies the known-fake label for 1,402 images. Type B withholds the label and combines these images with 1,140 real images, giving 2,542 images for detection evaluation. Four fake images have empty reference artifact sets. They remain in detection evaluation but are excluded from evidence scoring, leaving 1,398 eligible fake images.

Following TriDF (Jiang-Lin et al., 2026), the evidence metrics adapt Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018). Let P_i be the artifact set mapped from the model's explanation and G_i the reference artifact set. For each eligible image,

$$\mathrm{Cover}_i = \frac{|P_i \cap G_i|}{|G_i|}, \tag{15}$$

$$\mathrm{CHAIR}_i = \frac{|P_i \setminus G_i|}{|P_i|}, \tag{16}$$

$$\mathrm{Prec}_i = 1 - \mathrm{CHAIR}_i, \tag{17}$$

$$F_{0.5,i} = \frac{1.25\,\mathrm{Prec}_i \mathrm{Cover}_i}{0.25\,\mathrm{Prec}_i + \mathrm{Cover}_i}, \tag{18}$$

$$\mathrm{Hal}_i = \mathbf{1}[\mathrm{CHAIR}_i > 0]. \tag{19}$$

An empty reported artifact set receives Cover $= 0$, CHAIR $= 1$, Hal $= 1$, and $F_{0.5} = 0$. We also set $F_{0.5,i} = 0$ when its denominator is zero. Each reported metric is the mean of its per-image values over the eligible fake images.

These scores measure agreement with reference artifact sets. An unmatched observation can still be visually valid if it is absent from the reference; the FakeClue-Evidence protocol provides a separate assessment against the image.

C.3 FAKECLUE-EVIDENCE EVALUATION

On FakeClue (Wen et al., 2026), detection accuracy and macro-F1 use all 5,000 images. Evidence metrics use the 3,192 fake images. The model is not supplied with the authenticity label.

The claim evaluator receives one image–claim pair and judges visual-semantic support, pixel-forensic support, and forensic direction. Each support field has four states: supported, contradicted, uncertain, or not applicable. Direction indicates support for fake, support for real, or a neutral implication. A separate text-only call groups duplicate, paraphrased, contained, or artificially split observations. Every parsed claim belongs to exactly one group.

Claim states are assigned in the following order. A contradiction on either support axis gives H. Among the remaining claims, those with no supported axis receive U. Claims with visual support receive S when their forensic direction agrees with the label and $\mathrm{D}_{\mathrm{dir}}$ otherwise. Each semantic group takes the first state present under $\mathrm{H} \succ \mathrm{U} \succ \mathrm{D}_{\mathrm{dir}} \succ \mathrm{S}$.

This evaluation mapping differs from the training mapping in Appendix B.2. Evaluation state H denotes visual contradictions, while training state H also includes invalid and wrong-direction claims. The evaluator does not apply the training-only external-dependency and atomicity rules; claims requiring unavailable information follow its visual-support judgment.

For evaluation image i, let $\mathcal{S}_i$, $\mathcal{H}_i$, $\mathcal{U}_i$, and $\mathcal{D}_{\mathrm{dir},i}$ collect the groups assigned to each state. Their total number is $C_i = |\mathcal{S}_i| + |\mathcal{H}_i| + |\mathcal{U}_i| + |\mathcal{D}_{\mathrm{dir},i}|$. For $M = 3{,}192$ images, the metrics are

$$\mathrm{Evi@Img} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}[|\mathcal{S}_i| > 0], \tag{20}$$

$$\mathrm{Sup/Img} = \frac{1}{M}\sum_{i=1}^{M} |\mathcal{S}_i|, \tag{21}$$

$$\mathrm{Precision} = \frac{1}{M}\sum_{i=1}^{M} \frac{|\mathcal{S}_i|}{\max(C_i, 1)}, \tag{22}$$

$$\mathrm{Hall.\ Rate} = \frac{1}{M}\sum_{i=1}^{M} \frac{|\mathcal{H}_i|}{\max(C_i, 1)}. \tag{23}$$

Evi@Img measures how often an image receives supported evidence, while Sup/Img measures the number of supported evidence groups per image. Precision accounts for all four group states. Hallucination Rate (Hall. Rate) counts only visual contradictions in its numerator; uncertain and wrong-direction groups remain in its denominator.

Empty explanations and sole Notices contribute zero to both per-image fractions and remain in the full image denominator. A lower Hallucination Rate can therefore result from producing evidence on fewer images. We report Evi@Img and Sup/Img alongside precision and hallucination to distinguish this effect from an improvement in the reported evidence.

C.4 LOKI-Explain evaluation

LOKI-Explain (Ye et al., 2025) supplies the known-fake label for all 229 images. The evaluator receives the original image, the model's explanation, global and regional reference descriptions, and the ground-truth crops. It scores four axes: Global Quality, Completeness, Correctness, and Plausibility.

We average three evaluator calls per image, retaining every image regardless of the model's authenticity decision. Each call assigns an integer score from 0 to 5 on each axis. Let $\bar{s}_{ik}$ be the mean of the three scores for image i on axis k. The reported score is

$$Q_k = \frac{20}{M}\sum_{i=1}^{M} \bar{s}_{ik}, \qquad M = 229,$$

on a 0–100 scale. Overall is the mean of the four unrounded axis scores. We report TriDF, LOKI-Explain, and FakeClue-Evidence separately because their protocols and metric definitions differ.

D Additional Benchmark Results and Ablations

D.1 Additional detection results

Table 8 complements the main detection comparison with accuracy on the five primary benchmarks. Table 9 separately reports macro-F1 and balanced accuracy on Nano-banana-150K and fake recall on GPT-Image2-wild. The latter contains only fake images, so its recall is not combined with macro-F1 in the primary average.

Table 8: Detection accuracy (%) on the five primary benchmarks. Avg. is their unweighted mean.

Method	GenImage	LOKI-Det	SoFake-OOD	FakeBench	AIGIBench	Avg.↑
General-purpose MLLMs						
InternVL3.5-8B	66.21	56.25	52.11	77.65	64.62	63.37
InternVL3.5-38B	62.34	52.64	58.24	74.90	66.61	62.95
Qwen3-VL-8B-Instruct	74.44	66.04	78.63	84.70	76.72	76.11
Qwen3-VL-32B-Instruct	66.73	59.13	73.91	77.63	74.70	70.42
Binary detectors						
UniFD (CVPR'23)	65.34	46.96	49.94	67.62	60.83	58.14
ProDet (NeurIPS'24)	49.52	59.13	51.00	48.33	51.82	51.96
CO-SPY (CVPR'25)	78.00	59.95	63.41	85.62	71.52	71.70
D^3 (CVPR'25)	97.92	74.47	72.74	86.23	79.81	82.23
Effort (ICML'25)	76.60	68.06	60.34	71.75	61.74	67.70
Explanation-capable detectors						
FakeShield (ICLR'25)	87.44	62.61	80.71	88.73	65.71	77.04
SIDA (CVPR'25)	58.34	49.80	62.40	62.23	60.52	58.66
FakeVLM (NeurIPS'25)	**99.23**	**85.25**	75.41	94.23	79.61	86.75
MIRAGE-R1 (AAAI'26)	59.91	54.44	62.84	63.10	55.03	59.06
VERITAS (ICLR'26)	59.42	40.64	41.34	61.50	47.54	50.09
Ours						
CLOVER-SFT	<u>98.14</u>	<u>84.26</u>	<u>93.73</u>	**96.72**	89.73	**92.52**
CLOVER-RL (Stage I)	98.01	83.54	**93.84**	<u>96.45</u>	<u>90.63</u>	<u>92.49</u>
CLOVER-RL (Stage II)	97.93	83.18	93.70	96.10	**91.02**	92.39

D.2 Complete TriDF metrics

Table 10 separates reference coverage, CHAIR, the incidence of explanations containing unmatched artifacts, and $F_{0.5}$. From SFT through Stage II, reference coverage increases and CHAIR decreases under both prompt settings. The gains in $F_{0.5}$ therefore reflect both broader agreement with the references and fewer unmatched artifacts.

Type B detection accuracy decreases slightly as evidence scores improve. Evidence scoring retains all eligible fake images, including false negatives with empty reported artifact sets, under the protocol in Appendix C. The improvement therefore does not result from restricting evidence evaluation to correctly detected fakes.

D.3 Complete LOKI-Explain and FakeClue-Evidence metrics

Tables 11 and 12 show where explanation quality improves and where competing models remain stronger. On LOKI-Explain, Stage II exceeds Qwen3-VL-32B in Correctness, while Qwen3-VL-32B leads on the other axes and Overall. The improvement over CLOVER-SFT therefore does not imply leadership on every explanation criterion.

D.4 Supervision and coverage

Tables 13 and 14 provide the full explanation metrics for the supervision and evidence-coverage comparisons in Section 4.3. Sampling constraints, binary coverage configurations, and annotation reuse are specified in Appendix A.3. The additional corpus-size comparison below examines how much binary data to combine with the fixed evidence corpus. FC denotes FakeClue-Evidence. Det. Avg. consistently denotes mean macro-F1 across the five primary detection benchmarks.

Binary corpus size. With the evidence corpus fixed, increasing binary data from 50K to 100K raises detection macro-F1 from 90.21% to 92.44% with only a small loss in FakeClue-Evidence Precision

Table 9: Detection on additional benchmarks (%). BA denotes balanced accuracy. GPT-Image2-wild contains only fake images, so it reports fake recall.

Method	Nano-banana-150K		GPT-Image2-wild
	Macro-F1↑	BA↑	Fake recall↑
General-purpose MLLMs			
InternVL3.5-8B	50.32	54.73	75.63
InternVL3.5-38B	47.63	51.30	80.00
Qwen3-VL-8B-Instruct	52.42	63.40	77.40
Qwen3-VL-32B-Instruct	47.32	60.23	71.21
Binary detectors			
UniFD (CVPR'23)	33.36	49.44	0.31
ProDet (NeurIPS'24)	44.70	48.03	**90.53**
CO-SPY (CVPR'25)	45.70	51.43	26.14
D^3 (CVPR'25)	47.73	53.01	26.71
Effort (ICML'25)	43.61	43.63	74.70
Explanation-capable detectors			
FakeShield (ICLR'25)	40.80	52.73	37.30
SIDA (CVPR'25)	36.24	54.62	20.11
FakeVLM (NeurIPS'25)	68.51	68.81	59.13
MIRAGE-R1 (AAAI'26)	39.91	40.62	38.10
VERITAS (ICLR'26)	28.90	33.24	18.41
Ours			
CLOVER-SFT	**80.72**	83.83	<u>86.73</u>
CLOVER-RL (Stage I)	<u>80.53</u>	**84.32**	85.80
CLOVER-RL (Stage II)	80.12	<u>84.22</u>	85.33

(Table 13(b)). However, doubling the binary corpus again to 200K brings a smaller detection gain to 92.73%, while precision falls from 54.84 to 53.87. Thus, 100K binary images offer a useful balance between detection and explanation quality among the tested settings, with an evidence corpus roughly 10% its size.

This conclusion applies to the one-epoch setup. Larger corpora also receive more updates, so the comparison reflects both data volume and training budget (Appendix A.3).

Reusing evidence annotations. We reuse the same 4K annotations either for explainable detection or for label-conditioned explanation, keeping the number of training presentations fixed (Table 14(b)). Supplying the label raises FakeClue-Evidence Precision from 51.40 to 54.84 with little change in detection. The gain therefore comes from how the annotations are reused, rather than additional annotations or training presentations. This supports focusing the additional task on explaining a known decision instead of repeating the full detection task.

D.5 Judges, grouping, and reward schedules

Choice of claim judge. Table 15 extends the main comparison to InternVL3.5-38B. Using this judge yields 38.67 TriDF-B $F_{0.5}$ and 58.49 FakeClue-Evidence Precision, both below the results obtained with the fixed self-judge. These are downstream training outcomes, not direct measurements of judge accuracy. They do not establish superiority over untested RL judges such as Qwen3.8-Max, which is used only for evaluation here.

The *No claim judge* control treats every well-formed claim as provisionally supported while retaining semantic grouping. However, explanation precision falls and hallucinations increase under this control. Semantic grouping therefore complements image-based judgments but cannot replace them: distinct observations can still be unsupported.

Table 10: Complete TriDF results. Type A supplies the known-fake label; Type B withholds it. Both score evidence on 1,398 eligible fake images; Type B accuracy uses all 2,542 images. Cover measures reference coverage; CHAIR measures the unmatched-artifact fraction, and Hal is the fraction of explanations with CHAIR > 0.

(a) Type A: known fake

Method	Cover↑	CHAIR↓	Hal.↓	$F_{0.5}$ ↑
General-purpose MLLMs				
InternVL3.5-8B	37.78	57.66	94.71	39.33
InternVL3.5-38B	33.78	66.48	98.57	32.39
Qwen3-VL-8B-Instruct	46.69	66.58	99.43	34.63
Qwen3-VL-32B-Instruct	**66.42**	57.83	99.79	44.77
Explanation-capable detectors				
FakeShield (ICLR'25)	9.54	86.53	95.06	11.67
FakeVLM (NeurIPS'25)	35.73	79.18	99.79	22.33
MIRAGE-R1 (AAAI'26)	53.94	53.19	97.50	46.67
VERITAS (ICLR'26)	44.38	54.07	97.07	43.89
Ours				
CLOVER-SFT	40.26	52.55	83.33	43.47
CLOVER-RL (Stage I)	42.33	50.01	81.97	45.80
CLOVER-RL (Stage II)	43.60	**46.53**	**78.54**	**48.62**

(b) Type B: explainable detection

Method	Acc.↑	Cover↑	CHAIR↓	Hal.↓	$F_{0.5}$ ↑
General-purpose MLLMs					
InternVL3.5-8B	51.61	13.96	86.12	99.64	13.60
InternVL3.5-38B	51.85	7.06	92.69	99.79	7.08
Qwen3-VL-8B-Instruct	67.70	23.18	79.66	99.86	20.44
Qwen3-VL-32B-Instruct	70.22	**38.53**	74.01	100.00	27.40
Explanation-capable detectors					
FakeShield (ICLR'25)	47.05	1.52	98.10	99.71	1.76
FakeVLM (NeurIPS'25)	62.75	24.16	81.82	99.28	18.47
MIRAGE-R1 (AAAI'26)	71.16	27.37	72.57	98.50	26.69
VERITAS (ICLR'26)	71.28	23.44	72.28	97.78	25.87
Ours					
CLOVER-SFT	**81.39**	22.40	53.22	62.09	36.15
CLOVER-RL (Stage I)	81.31	23.70	49.96	57.80	38.65
CLOVER-RL (Stage II)	81.20	25.16	**46.88**	**53.72**	**41.01**

Semantic grouping. Table 16 compares counting every claim, exact-string deduplication, and semantic grouping under matched Stage II training budgets. The duplicate rate falls from 23.33 with claim-wise rewards to 11.05 with exact-string deduplication and 4.52 with semantic grouping. FakeClue-Evidence Precision, TriDF $F_{0.5}$, and LOKI-Explain Overall improve while detection remains similar. Unlike exact matching, semantic grouping counts paraphrases of the same observation only once. These results support grouping by meaning to reduce redundant evidence and improve precision beyond exact-string deduplication.

Reward identity and order. Table 17(a) compares the coverage and faithfulness rewards after 100 updates from the same SFT checkpoint. The coverage reward produces supported evidence on more images. In contrast, the faithfulness reward yields fewer hallucinations but lower supported-image coverage. The two rewards therefore favor different balances between evidence discovery and the exclusion of unsupported claims.

Table 11: Complete LOKI-Explain results on 229 known-fake images. Scores range from 0 to 100; Overall is the mean of the four axes.

Method	Global Quality↑	Completeness↑	Correctness↑	Plausibility↑	Overall↑
General-purpose MLLMs					
InternVL3.5-8B	57.03	53.54	47.48	65.04	55.77
InternVL3.5-38B	55.49	49.11	43.67	66.64	53.73
Qwen3-VL-8B-Instruct	72.63	72.58	62.71	73.97	70.47
Qwen3-VL-32B-Instruct	**74.21**	**78.78**	66.93	**87.22**	**76.78**
Explanation-capable detectors					
FakeShield (ICLR'25)	42.68	26.67	27.69	49.75	36.70
FakeVLM (NeurIPS'25)	62.01	55.02	55.69	63.52	59.06
MIRAGE-R1 (AAAI'26)	54.29	52.58	48.33	66.46	55.41
VERITAS (ICLR'26)	53.54	49.46	46.40	62.18	52.90
Ours					
CLOVER-SFT	66.58	67.83	62.39	73.94	67.69
CLOVER-RL (Stage I)	69.72	74.91	64.75	72.20	70.39
CLOVER-RL (Stage II)	73.42	74.53	**67.51**	78.60	73.52

Table 12: Complete FakeClue-Evidence results. Detection uses all 5,000 images; evidence scores use the 3,192 fake images. Scores range from 0 to 100, except Sup/Img, the mean number of supported evidence groups per image.

Method	Acc.↑	Macro-F1↑	Evi@Img↑	Sup/Img↑	Precision↑	Hall. Rate↓
General-purpose MLLMs						
InternVL3.5-8B	50.14	48.85	26.03	0.7835	14.42	12.18
InternVL3.5-38B	57.12	57.08	43.36	1.1181	29.92	9.39
Qwen3-VL-8B-Instruct	54.28	52.98	28.82	1.2428	22.22	11.74
Qwen3-VL-32B-Instruct	56.34	55.44	32.61	1.6093	25.52	12.08
Explanation-capable detectors						
FakeShield (ICLR'25)	38.02	29.80	3.45	0.0805	1.80	12.41
FakeVLM (NeurIPS'25)	**98.96**	**98.88**	**74.87**	1.6582	34.33	25.45
MIRAGE-R1 (AAAI'26)	44.26	39.73	13.19	0.4380	7.77	15.10
VERITAS (ICLR'26)	72.70	72.55	47.53	1.0075	27.79	12.97
Ours						
CLOVER-SFT	90.94	90.10	65.48	1.4756	54.84	14.55
CLOVER-RL (Stage I)	90.54	89.79	70.43	1.6338	59.05	13.42
CLOVER-RL (Stage II)	90.50	89.73	70.43	**1.7660**	**61.46**	**9.29**

Panel (b) compares five schedules with 200 accepted RL updates each and includes an Extra SFT control. Coverage followed by faithfulness achieves the best TriDF-B $F_{0.5}$ and FakeClue-Evidence Precision among these schedules. Faithfulness-only training minimizes hallucinations but yields the lowest supported-image coverage. Relative to the proposed order, reversing the stages increases coverage but also hallucinations, whereas a constant mixed reward gives similar coverage and lower precision.

Extra SFT leaves evidence quality close to the supervised baseline despite matching the nominal update count. Thus, additional supervised updates do not reproduce the explanation gains from evidence RL under this control. The continuation protocol and its difference in computation are specified in Appendix A.3. Appendix B.5 explains why stronger penalties in the raw reward alone do not establish a benefit after GSPO normalization.

Table 13: Supervision mixture and binary corpus size. (a) SFT updates are matched by replaying the remaining data. The 15K evidence presentations use 11K annotated images. (b) The evidence corpus and binary-source proportions remain fixed. Each model trains for one epoch, so larger binary corpora receive more updates.

(a) Supervision combination

Condition	Binary (K)	Evidence (K)	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	FakeClue-Evidence Evi@Img↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
Binary only	100	0	**93.12**	22.86	15.66	50.31	31.76	23.99
Evidence only	0	15	71.18	42.15	28.76	62.81	49.33	16.41
Binary + Evidence	100	15	92.44	**43.47**	**36.15**	**65.48**	**54.84**	**14.55**

(b) Binary corpus size under one-epoch training

Binary SFT	Detection (%) Avg. Acc.↑	Detection (%) Det. Avg.↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Evi@Img↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
50K	90.24	90.21	**43.47**	35.39	67.68	**65.76**	**54.99**	**14.40**
100K (primary)	92.52	92.44	**43.47**	**36.15**	**67.69**	65.48	54.84	14.55
200K	**92.81**	**92.73**	42.25	35.78	67.35	65.16	53.87	15.30

Table 14: Evidence coverage and annotation reuse with 100K binary images. (a) Scene and generator coverage vary at a fixed budget of 11K annotated images and 15K evidence presentations. Counts denote scene domains and generator groups. (b) The same 4K annotations are reused for either explainable detection or label-conditioned explanation, with total presentations fixed.

(a) Evidence scenes and generators

Scenes	Generators	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Evi@Img↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
2	4	**92.44**	36.65	29.87	63.71	61.37	47.64	15.61
4	4	92.38	40.62	33.54	65.58	63.50	51.29	14.91
7	4	**92.44**	43.47	36.15	67.69	65.48	54.84	**14.55**
7	20	92.41	**43.69**	**36.30**	**67.73**	**65.66**	**54.98**	**14.55**

(b) Label-conditioned explanation versus explainable detection replay

Condition	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Evi@Img↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
Explainable detection replay	92.41	39.94	33.19	66.01	63.66	51.40	15.39
Label-conditioned explanation	**92.44**	**43.47**	**36.15**	**67.69**	**65.48**	**54.84**	**14.55**

Detection and evidence rewards. The controls in Table 18(a) remove one reward component while retaining the full task mixture. *No binary reward* assigns zero to parsed binary responses and -1 to unparseable ones. *No evidence reward* assigns zero after the evidence response passes the decision and format checks, retaining the failure penalties.

Table 15: Choice of claim judge under matched RL settings. The dynamic self-judge copies the Actor before each update; all other judges remain fixed. The control without a claim judge credits well-formed claims without image checks. Param. gives the claim judge's size relative to 8B.

Claim judge	Fixed in RL	Param. ratio	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
No claim judge	–	0.0×	92.27	36.82	33.68	66.75	52.19	20.25
Dynamic self-judge	No	1.0×	92.32	43.42	37.47	69.72	56.16	12.68
Qwen3-VL-8B	Yes	1.0×	92.33	44.11	37.80	70.29	57.17	11.96
Qwen3-VL-32B	Yes	4.0×	92.28	46.23	39.45	70.77	58.09	11.04
InternVL3.5-38B	Yes	≈ 4.8×	92.34	45.06	38.67	71.07	58.49	11.36
Self-judge (ours)	Yes	1.0×	**92.35**	**48.62**	**41.01**	**73.52**	**61.46**	**9.29**

Table 16: Effect of evidence grouping at matched Stage II training budgets. Duplicate Rate is the percentage of repeated or overlapping claims, computed from the mean of $(N_i - C_i)/\max(N_i, 1)$ over 3,192 FakeClue fake images. Here N_i and C_i are claim and group counts; a common text-only evaluator supplies the groups.

Accounting	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓	Duplicate rate↓
Claim-wise	92.33	43.87	38.40	70.52	56.73	13.45	23.33
Exact-string deduplication	92.28	45.51	39.24	71.88	59.03	11.45	11.05
Semantic grouping	**92.35**	**48.62**	**41.01**	**73.52**	**61.46**	**9.29**	**4.52**

Table 17: Reward choice and training order from the same SFT checkpoint. (a) Coverage and faithfulness rewards after 100 updates. (b) RL schedules after 200 updates with matched training examples and order; Extra SFT uses 200 additional supervised updates. Update counts do not imply equal computation. Cov./Faith. denote coverage/faithfulness.

(a) Early reward behavior: 100 updates

Schedule	Updates	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Evi@Img↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
Coverage only	100	92.42	**45.80**	**38.65**	**70.39**	**70.43**	**59.05**	13.42
Faithfulness only	100	**92.43**	43.35	31.33	68.49	66.48	54.34	**7.92**

(b) Final reward comparison: 200 updates

Schedule	Det. Avg. (%)↑	TriDF $F_{0.5}$ ↑ A↑	TriDF $F_{0.5}$ ↑ B↑	LOKI Overall↑	FakeClue-Evidence Evi@Img↑	FakeClue-Evidence Sup/Img↑	FakeClue-Evidence Precision↑	FakeClue-Evidence Hall. Rate↓
Coverage only	92.38	45.61	38.86	70.97	**73.68**	1.6895	59.50	13.96
Faithfulness only	92.41	41.62	26.92	68.79	65.19	1.4527	53.18	**5.21**
Mixed reward	92.32	45.21	38.77	71.38	69.83	1.6977	59.03	11.35
Faith. → Cov.	92.28	45.12	39.18	71.57	71.43	1.7340	59.41	12.68
Cov. → Faith.	92.35	**48.62**	**41.01**	**73.52**	70.43	**1.7660**	**61.46**	9.29
Extra SFT	**92.63**	43.82	36.58	67.92	66.04	1.4890	55.23	14.30

Without evidence rewards, explanation quality remains close to the SFT baseline despite continued RL. Removing binary rewards instead improves FakeClue-Evidence Precision, but detection macro-F1 falls to 90.77% and TriDF-B $F_{0.5}$ also decreases relative to the complete system. The components therefore serve complementary purposes: evidence rewards improve explanations, while binary rewards help preserve detection as evidence quality is optimized.

The three-task mixture in RL. Table 18(b) compares prompt counts of $0/64/0$, $16/48/0$, and $16/32/16$ per update for binary detection, explainable detection, and label-conditioned explanation, respectively. All variants use 200 accepted updates. The last comparison keeps binary exposure fixed and replaces explainable detection prompts with label-conditioned explanation prompts.

Adding binary prompts to explainable detection raises detection macro-F1 from 87.63% to 92.34%, although FakeClue-Evidence Precision decreases. Dedicated binary prompts thus help preserve detection even when explanation responses already penalize incorrect decisions, but the two-task mixture does not improve both capabilities.

At the same binary exposure, introducing label-conditioned explanation raises FakeClue-Evidence Precision from 59.47 to 61.46 with little change in detection. Together, these comparisons support the three-task mixture: binary detection preserves authenticity discrimination, while label-conditioned explanation improves evidence quality by focusing on a known decision.

Table 18: Reward components and task mixtures under the two-stage, 200-update schedule. (a) Reward controls retain the full task mixture and response checks. (b) The three rows use $0/64/0$, $16/48/0$, and $16/32/16$ prompts per update for binary detection, explainable detection, and label-conditioned explanation, respectively.

(a) Reward-component controls

Reward configuration	Det. Avg.	TriDF $F_{0.5}$ ↑		LOKI	FakeClue-Evidence		
	(%)↑	A↑	B↑	Overall↑	Evi@Img↑	Precision↑	Hall. Rate↓
No binary reward	90.77	**48.96**	39.76	**74.18**	**72.71**	**63.14**	**8.94**
No evidence reward	**92.57**	43.17	36.26	67.18	66.32	55.41	14.57
Complete system	92.35	48.62	**41.01**	73.52	70.43	61.46	9.29

(b) RL prompt mixtures

RL prompt mixture	Det. Avg.	TriDF $F_{0.5}$ ↑		LOKI	FakeClue-Evidence		
	(%)↑	A↑	B↑	Overall↑	Evi@Img↑	Precision↑	Hall. Rate↓
Explainable detection only	87.63	48.58	39.11	73.29	70.05	60.78	10.16
Binary + explainable detection	92.34	45.11	38.49	72.13	69.58	59.47	10.06
All three tasks	**92.35**	**48.62**	**41.01**	**73.52**	**70.43**	**61.46**	**9.29**

E Evidence Analysis and Source Adaptation

E.1 Independent expert assessment

We use independent image-forensics experts to examine both the self-judge's judgments and the quality of the generated explanations. The scoring form in Table 19 separates visual support from forensic direction.

Agreement with expert consensus. The agreement study uses 600 claims from 300 images. The self-judge and held-out experts are compared against an independent expert consensus. Agreement

Table 19: Scoring form for independent image-forensics experts. Each claim receives one visual-support rating and one forensic-direction rating.

Criterion	Rating	Assessment
Visual support	Supported	The described observation is supported by the image.
	Uncertain	The image does not allow a reliable decision about the observation.
	Hallucinated	The described observation is contradicted by the image.
Forensic direction	Aligned	The observation's authenticity implication agrees with the ground-truth label.
	Nonaligned	The implication is neutral or conflicts with the ground-truth label.

and Cohen's κ are reported separately for each axis, and the expert comparison averages individual scores.

On this sample, the self-judge has higher agreement on visual support and slightly lower agreement on forensic direction than the mean held-out expert (Table 20(a)). This comparison provides a separate check on the self-judge's outputs under our protocol; it does not establish superiority over human experts or other judges more generally.

Paired explanation quality. The paired study compares SFT, Stage I, and Stage II on the same 300 fake images with fixed decoding. Experts receive responses in randomized order without model identity or automated judgments. Consensus labels and the grouping rules in Appendix C.3 determine the metrics. All 300 images remain in the calculation; empty responses and sole Notices contribute zero.

Precision increases from 53.15 at SFT to 60.26 at Stage II, while Hallucination Rate decreases from 14.16 to 10.68. Evi@Img rises from 64.67 to 70.33 at Stage I and then decreases to 69.33 at Stage II. These independent ratings support improved evidence reliability, while showing that stronger selection can leave fewer images with supported evidence than at Stage I.

Table 20: Independent expert assessment. (a) The self-judge and held-out experts are compared with an independent expert consensus on 600 claims from 300 images. Expert scores average individual agreement and Cohen's κ. (b) Blinded ratings compare evidence quality across stages on the same 300 fake images, including empty responses and Notices.

(a) Agreement against a shared independent consensus

Comparison	Visual support		Forensic direction	
	Agreement (%)	κ	Agreement (%)	κ
Held-out experts (mean)	<u>90.80</u>	<u>0.850</u>	**93.80**	**0.821**
Self-judge	**92.80**	**0.883**	<u>92.70</u>	<u>0.783</u>

(b) Paired expert assessment of evidence quality

Stage	Evi@Img	Precision	Hall. Rate↓
SFT	64.67	53.15	14.16
Stage I	**70.33**	<u>57.21</u>	<u>12.06</u>
Stage II	<u>69.33</u>	**60.26**	**10.68**

E.2 Adapting to new sources without new evidence annotations

This experiment tests whether binary supervision can extend detection to new generators while the evidence corpus remains fixed. All four conditions in Table 21 start from the same checkpoint. The new-source generator families occur in neither initial training corpus.

Table 21: Source adaptation from a shared checkpoint. Adapted models share the update and presentation budgets. The annotation condition reallocates 10K evidence presentations from old to new sources. Detection reports macro-F1; evidence metrics follow the FakeClue-Evidence protocol. Scores range from 0 to 100 and are ranked separately within each test pool.

Adaptation arm	**Pool**	**Updates**	**New ann.**	**Macro-F1↑**	**Evidence**		
					Evi@Img↑	**Precision↑**	**Hall. Rate↓**
Base (no adapt.)	old	0	0	91.20	**66.70**	**55.30**	15.00
Base (no adapt.)	new	0	0	76.80	52.50	40.40	16.80
+ old binary	old	200	0	<u>92.00</u>	64.90	54.30	**13.70**
+ old binary	new	200	0	75.80	53.90	<u>42.70</u>	<u>15.60</u>
+ new binary	old	200	0	91.40	<u>66.00</u>	<u>54.90</u>	<u>14.40</u>
+ new binary	new	200	0	<u>85.30</u>	<u>54.50</u>	42.20	16.10
+ binary + ann. realloc.	old	200	10000	**93.00**	63.60	52.50	15.20
+ binary + ann. realloc.	new	200	10000	**85.60**	**65.50**	**53.20**	**13.10**

The first three conditions compare no adaptation, old-source binary adaptation, and new-source binary adaptation. Both adapted conditions retain the original 11K evidence images, targets, and 4K label-conditioned subset. They use 200 updates with 200 binary and 75 evidence presentations per update, giving 40K binary presentations and 15K old-evidence presentations without new annotations.

The fourth condition uses the same new-source binary data but replaces 10K of the 15K old-evidence presentations with annotated new-source presentations. All adapted conditions share the update and presentation budgets, class balance, preprocessing, decoding, and checkpoint selection. Old-source and new-source test pools are evaluated separately. Because the added sources occur in adaptation training, this experiment measures adaptation rather than generalization to sources that remain unseen.

With the evidence corpus fixed, adaptation using new-source binary images raises detection macro-F1 on the new sources from 76.80 to 85.30. In contrast, equally many updates using old-source images yield 75.80 (Table 21). The improvement therefore depends on exposure to the new sources, rather than additional updates alone. Since old-source evidence scores change little, this comparison supports expanding detection to new generators without collecting new explanation annotations.

Adding new-source evidence annotations further raises new-source precision from 42.20 to 53.20, but old-source precision falls from 54.90 to 52.50 as its evidence presentations are replaced. Thus, new-source annotations can further improve explanations for the added sources, but replacing old-source evidence presentations creates a tradeoff between the two source groups.

E.3 Low-level evidence and abstention

Table 22 examines whether allowing the qualified low-level evidence (QLE) defined in Appendix A.2 and providing the Notice option improve explanation quality.

Allowing low-level evidence. The VSE-only control retains the same source images, presentation counts, pretrained initialization, and update budgets. It removes QLE from SFT targets: mixed targets retain their VSE claims, while QLE-only targets become Notice responses. Its prompts request semantic observations, and RL treats QLE as ineligible. It therefore uses its own SFT checkpoint and matching self-judge. Evaluation retains the same broad evidence scope for all variants, and decoding remains unrestricted.

Allowing QLE raises Evi@Img from 63.97 to 70.43 and FakeClue-Evidence Precision from 55.42 to 61.46. However, Hallucination Rate also rises from 6.53 to 9.29. Allowing low-level cues therefore expands supported evidence coverage and improves precision, but at the cost of more hallucinated claims. This conclusion concerns the evidence policy across SFT and RL, rather than adding low-level wording at inference alone.

Allowing evidence abstention. The Notice-penalty control penalizes an otherwise valid sole Notice as an invalid response. It retains the full model's SFT initialization, data, prompts, judges, other format checks, decision penalties, reward schedule, and update budget. Decoding remains unrestricted, and evaluation uses the same rules for both models.

Although disabling Notice slightly raises supported-image coverage, it lowers TriDF-B $F_{0.5}$ and FakeClue-Evidence Precision. It also increases Hallucination Rate from 9.29 to 17.95. These results support allowing abstention: removing the option produces supported evidence on slightly more images, but makes the reported evidence less reliable.

Table 22: Evidence scope and abstention under matched image and update budgets. VSE and QLE denote visible-semantic and qualified low-level evidence. VSE-only excludes QLE throughout SFT and RL; the Notice-penalty control changes only the RL reward for an otherwise valid Notice to -1.

Training configuration	**TriDF-B**	**LOKI**	**FakeClue-Evidence**			**Notice**
	$F_{0.5}$ ↑	**Overall↑**	**Evi@Img↑**	**Precision↑**	**Hall. Rate↓**	**rate**
VSE only + Notice	37.46	<u>72.32</u>	63.97	55.42	**6.53**	25.34
VSE + QLE, Notice penalty	<u>38.84</u>	71.86	**70.96**	<u>57.89</u>	17.95	0.00
Full model (VSE + QLE + Notice)	**41.01**	**73.52**	<u>70.43</u>	**61.46**	<u>9.29</u>	17.89

Evidence availability when the model abstains. Image-forensics experts independently assess the original Notice responses using the four categories in Table 23. These categories describe visible evidence availability rather than the model's internal reason for abstaining.

Usable but overlooked evidence accounts for 13.13% of Notices and 2.35% of all eligible fake images. This assessment identifies the cost of withholding an explanation and distinguishes images without usable cues from those on which further evidence discovery remains possible.

Table 23: Independent expert assessment of evidence availability in images receiving a Notice.

Availability category	**Images**	**Share of Notices**
Insufficient visible cues	302	52.89%
Ambiguous forensic direction	132	23.12%
Only out-of-scope candidate explanations	62	10.86%
Usable evidence overlooked	75	13.13%

E.4 Qualitative comparisons across training stages

Figures 4–7 compare SFT and both RL stages on four additional images. Each figure preserves the complete English evidence statements. The assistant markers, leading ellipses, highlights, and footer notes are presentation annotations. Lettered boxes mark every enlarged region, with matching letters and colors in the detail views. All crops come directly from the input image. The real-image cases illustrate explanation behavior separately from the fake-image explanation metrics. The final case complements the abstention analysis in Appendix E.3.

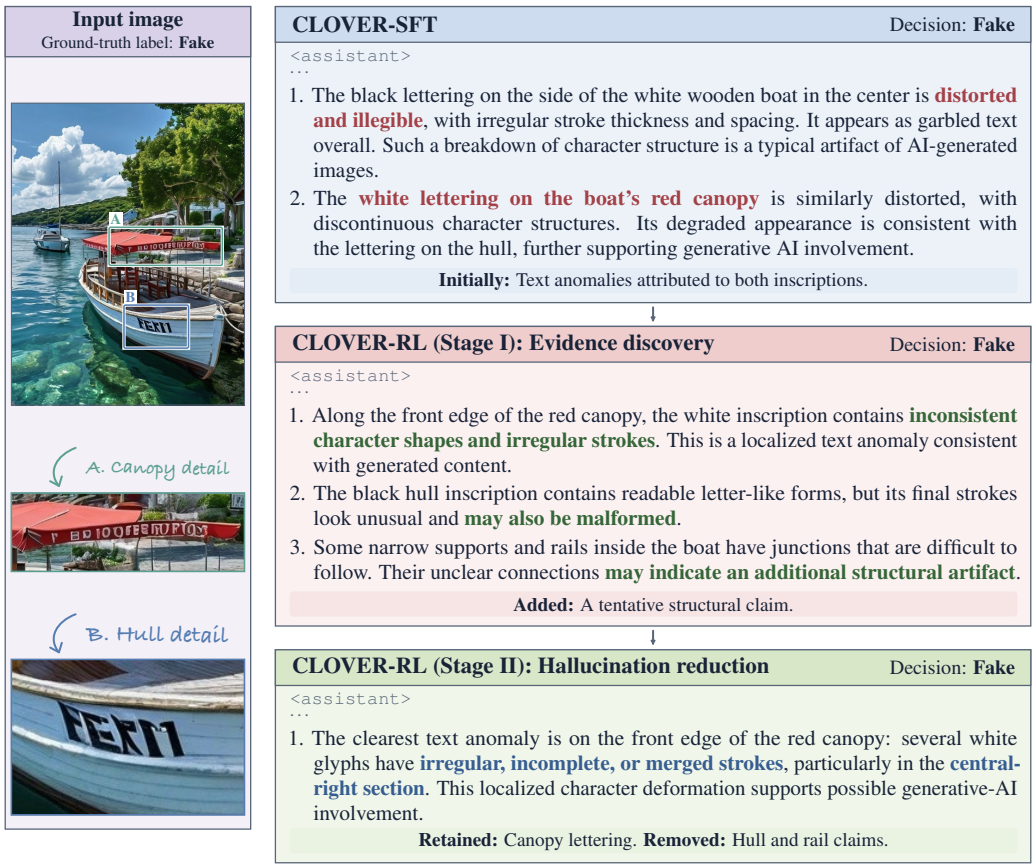
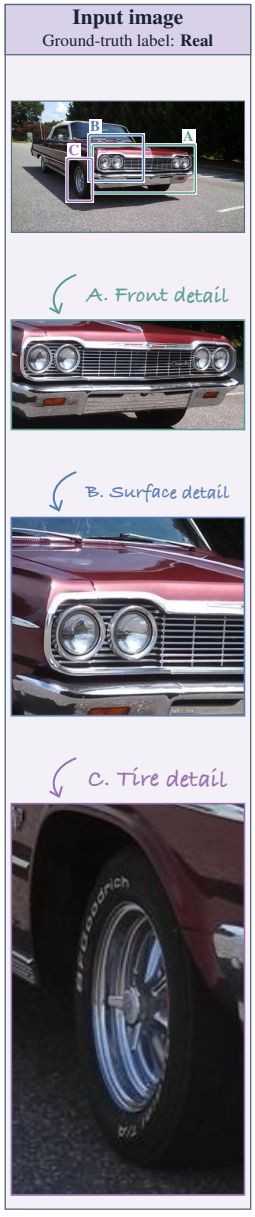


Figure 4: **More precise evidence localization.** Stage I localizes irregular lettering on the canopy but also raises tentative concerns about the hull inscription and rail joints. Stage II retains the canopy evidence and removes these less certain claims.



CLOVER-SFT Decision: **Real**

<assistant>
...

1. The lighting across the car, curb, tree line, and ground is consistent. Highlights on the front of the car, **windshield reflections**, and highlights on the chrome trim all agree with strong sunlight. The shadow beneath the car also follows the overall lighting direction. This coherent, natural lighting is consistent with a real photograph.
2. The perspective relationships among the wheels, fenders, bumper, grille, and headlights are coherent. There are no obvious breaks, misalignments, or interpenetrations between these components, and the overall structure matches the geometry of a real vehicle.
3. The paintwork shows natural highlight gradients and reflections in strong light. **Reflections on the chrome bumper and grille** are continuous, with natural transitions consistent with real metal surfaces in outdoor sunlight.
4. The trees, shrubs, and distant objects in the background exhibit natural depth-of-field blur relative to the sharp foreground car, consistent with camera imaging.

Initially: Broad cues with overlapping reflection descriptions.

↓

CLOVER-RL (Stage I): Evidence discovery Decision: **Real**

<assistant>
...

1. The **horizontal grille bars remain aligned** between the headlights, and the bumper and fender edges follow the same perspective. These connected parts form a coherent vehicle structure.
2. The visible tires meet the road, and the shadow extends continuously beneath and beside the car. The contact and shadow positions are consistent with its placement on the asphalt.
3. The chrome bumper contains sharp highlights and darker reflections, while the painted hood has broader highlight gradients. These different responses are consistent with the two surface finishes.
4. The **white lettering on the visible tire** follows the curvature of its sidewall without an abrupt change in orientation, providing an additional localized observation.
5. The **chrome trim and grille also reflect** the bright surroundings coherently, further supporting natural illumination.

Added: Tire lettering and more localized observations.

↓

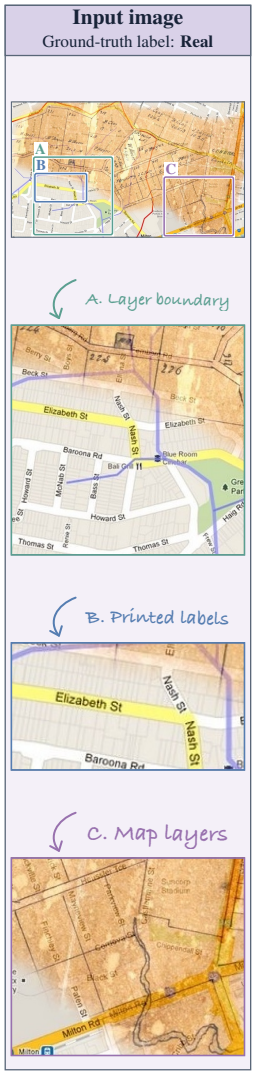
CLOVER-RL (Stage II): Hallucination reduction Decision: **Real**

<assistant>
...

1. The grille bars, circular headlights, bumper, and fender edges form **continuous, mutually consistent geometry** without visible breaks at their junctions.
2. The visible tires contact the asphalt, and the continuous shadow beneath the car agrees with its position and the direction of sunlight.
3. Sharp highlights and darker reflections on the chrome contrast naturally with the broader gradients on the painted body, giving a coherent account of the **different surface materials**.
4. The **white tire lettering** follows the sidewall's curvature with a consistent orientation. This localized printed detail is compatible with a photographed physical tire.

Retained: Local details. **Removed:** Repeated reflection claim.

Figure 5: **Refining evidence for a real image.** Stage I adds tire-lettering evidence and more localized observations of vehicle structure and materials. Stage II retains these details while removing an overlapping reflection claim.



CLOVER-SFT Decision: **Real**

<assistant>
...

1. Roads, parcel boundaries, building numbers, and place names are overlaid on the map in different orientations, sizes, and colors. Their overlaps with map lines and colored regions are broadly coherent, with no floating elements, breaks, or obvious misalignments, consistent with common annotation overlays on real maps.
2. The entire map has a **uniform aged-paper appearance**, with evenly distributed yellowing, mottling, grain, and slight fading. Street lines, parcel boundaries, and text all show a **similar degree of paper aging**. This consistent aged appearance is compatible with a scan or photograph of a real historical map.

Initially: The whole map is described as uniformly aged.

↓

CLOVER-RL (Stage I): Evidence discovery Decision: **Real**

<assistant>
...

1. The image contains **two distinct map styles**: handwritten parcel annotations on yellowed paper and a cleaner modern street map with printed labels and colored roads. Their coexistence is consistent with a map overlay.
2. Modern labels such as "Elizabeth St" and "Milton Rd" remain legible alongside the older handwritten names and numbers. Each layer retains its own lettering style.
3. The brown paper texture fades near the lower-left part of the overlay, **revealing the white modern map beneath it**. This transition is consistent with a translucent image layer.
4. The old and modern roads appear to have been **aligned precisely**, suggesting an accurate historical map registration.

Added: Two map layers; precise alignment is also claimed.

↓

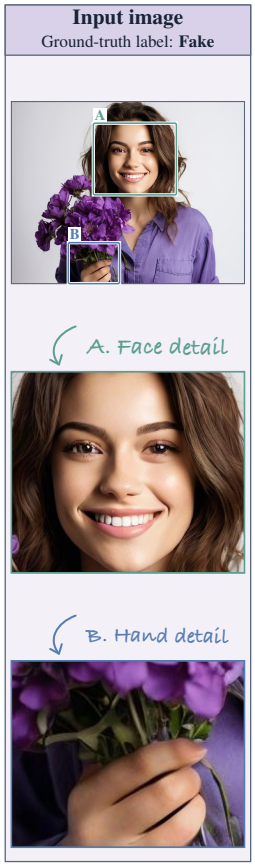
CLOVER-RL (Stage II): Hallucination reduction Decision: **Real**

<assistant>
...

1. A yellowed map with handwritten parcel notes overlaps a cleaner modern street map. The distinct text styles and road colors show that the image is **not a uniformly aged historical scan**.
2. Near the lower-left boundary of the brown overlay, modern streets and labels remain visible through or outside the older map layer. This is consistent with **conventional translucent compositing**.
3. Printed labels such as "Elizabeth St" and "Milton Rd" remain legible, while the older names and numbers follow a separate handwritten style. These locally coherent details support a layered map composition without requiring a generative explanation.

Retained: Layered structure. **Removed:** Precise alignment.

Figure 6: **More precise scene interpretation.** Stage I distinguishes the historical overlay from the modern map. Stage II retains evidence of their layered structure and removes the claim of precise alignment, yielding a more restrained explanation.



CLOVER-SFT Decision: **Fake**

<assistant>
...

1. The facial skin, especially on the forehead, cheeks, and bridge of the nose, has an **unusually uniform, smooth appearance**, with almost no visible pores, minor imperfections, or natural skin texture. The shading transitions are also excessively soft. This overall smoothness is more consistent with generative AI editing of facial attributes.
2. The hand holding the flowers has **structural abnormalities**: the finger lengths, joint proportions, and gripping pose look unnatural, while the boundaries between fingers are blurred and fused. These hand-structure irregularities are more consistent with generative AI generation or editing.

Initially: Strong claims about smooth skin and hand shape.

CLOVER-RL (Stage I): Evidence discovery Decision: **Fake**

<assistant>
...

1. The facial skin is much smoother than the sharply defined hair and eyelashes. This difference **may suggest** localized smoothing or synthetic treatment of the face.
2. The fingers around the flower stems overlap tightly, and some of their boundaries are difficult to trace. This **may indicate** a structural inconsistency in the grip.
3. Several fine boundaries within the overlapping bouquet are soft or indistinct. This **could be another sign** of synthetic detail, although the dense overlap **limits how confidently it can be interpreted**.

Added: Bouquet details, with explicit uncertainty.

CLOVER-RL (Stage II): Hallucination reduction Decision: **Fake**

<assistant>
...

1. The model's overall internal representation is more consistent with generative-AI participation, but no visual or machine-internal anomaly could be stably localized or independently verified, so this conclusion has **low explainability**.

Notice: Evidence abstention; the fake decision remains.

Figure 7: **Evidence abstention.** Stage II replaces tentative claims about skin, hand structure, and bouquet details with Notice. The fake decision is preserved, but the model abstains from providing visual evidence.