VISUAL THOUGHT MATTERS: LEARNING TO GROUND AND REFINE CONTINUOUS REASONING FOR GENERALIZABLE DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable deepfake detection requires reasoning over subtle visual evidence that is difficult to faithfully express in discrete language. While multimodal large language models (MLLMs) can generate detailed reasoning chains, text-dominated reasoning may overlook fine-grained forensic cues. To this end, we propose COVE, a framework that integrates continuous forensic states into the reasoning trajectory to capture and leverage fine-grained visual evidence throughout reasoning. Complementary frozen visual experts supervise these states to encode appearance, semantic, and boundary cues, equipping the MLLM with fine-grained forensic perception. To encourage the model to use this evidence for decision making, we introduce a Counterfactual Token Utility Reward (CTUR), which quantifies the contribution of forensic states to the correct verdict through counterfactual suppression and optimizes their decision utility via reinforcement learning. We further introduce a contrastive grounding objective that contrasts the same target reasoning under semantically matched visual conditions and optimizes its grounding gain relative to a frozen reference model. This encourages the refined reasoning to depend on its supporting visual evidence, while complete-chain supervision and replay preserve reliable reasoning behavior. At inference, COVE uses a single MLLM to generate the complete forensic reasoning trajectory and authenticity verdict. Experiments on HydraFake demonstrate strong generalization across diverse out-of-distribution scenarios, particularly on unseen forgeries and data domains, while producing high-quality forensic reasoning. Together, these results demonstrate the effectiveness of COVE in connecting fine-grained visual representation, decision utility, and evidence-grounded reasoning for generalizable deepfake detection.

1 INTRODUCTION

Realistic face editing and image synthesis have made deepfake detection fundamentally a problem of generalization across generators, manipulation types, and image sources. Existing detectors exploit local artifacts and pretrained visual representations (Shiohara & Yamasaki, 2022; Tan et al., 2024; Ojha et al., 2023), yet the evidence indicative of manipulation can vary substantially across these settings. A manipulated face may exhibit irregular boundary transitions, whereas a synthesized image may contain inconsistent textures or facial structures. Reliable detection therefore requires capturing diverse, fine-grained forensic cues while reasoning about them in the broader visual context.

Multimodal large language models (MLLMs) provide a natural framework for organizing such evidence into explicit forensic reasoning. FakeVLM (Wen et al., 2025) learns artifact explanations, Veritas (Tan et al., 2026) trains structured forensic reasoning, and VIGIL (Li et al., 2026) performs regional inspection and injects the resulting visual evidence into its reasoning process. These advances extend deepfake detection beyond authenticity prediction toward reasoning about the evidence behind a decision. However, they also expose a fundamental challenge: ***how can an MLLM represent fine-grained forensic evidence during reasoning, actually use that evidence for its decision, and keep the resulting reasoning grounded in the image?***

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.

Two challenges arise when incorporating continuous visual evidence into forensic reasoning. First, subtle forensic cues are difficult to faithfully express in discrete language, motivating continuous visual states that preserve fine-grained perceptual information (Qin et al., 2025). Yet learning what these states encode does not ensure that the model actually uses them: representing forensic evidence and relying on it for the final decision are distinct objectives. Effective forensic reasoning therefore requires both learning informative continuous states and explicitly promoting their decision utility. Second, refining a reasoning trajectory does not necessarily strengthen its visual grounding. Supervision can make a target explanation more likely even when the input image does not support it. Reasoning refinement should instead increase support for an analysis specifically when its underlying visual evidence is present. This motivates contrasting the same reasoning under controlled visual conditions, so that refinement strengthens its dependence on the supporting image evidence (Yu et al., 2026).

To address these challenges, we propose COVE, a general deepfake detection framework that integrates visual forensic evidence into the LLM reasoning trajectory for evidence-aware and visually grounded reasoning. First, COVE integrates continuous forensic states into the reasoning trajectory and supervises them with complementary frozen visual experts, equipping the MLLM to capture fine-grained appearance, semantic, and boundary cues throughout reasoning. Second, we introduce the Counterfactual Token Utility Reward (CTUR), which counterfactually suppresses forensic-state inputs and measures the resulting change in support for the correct verdict. Optimizing this signal through reinforcement learning explicitly encourages the model to use the learned forensic evidence for decision making. Third, we introduce Contrastive Forensic Grounding to refine complete reasoning trajectories. For the same target reasoning, we contrast its support under semantically matched visual conditions and optimize the resulting grounding gain relative to a frozen reference model. Together, these components establish a progressive connection from fine-grained forensic representation to decision utility and ultimately to evidence-grounded reasoning. Together, these components establish a progressive connection from fine-grained forensic representation to decision utility and evidence-grounded reasoning. Extensive experiments on HydraFake demonstrate state-of-the-art generalization across diverse evaluation settings, achieving 94.0% overall accuracy. Comprehensive ablations and analyses further validate the effectiveness of our key designs and the quality and robustness of the resulting forensic reasoning.

Our contributions are as follows:

- We introduce continuous forensic states into the reasoning trajectory and ground them with complementary visual experts, enabling MLLMs to capture fine-grained forensic evidence beyond discrete textual descriptions.
- We introduce CTUR, a counterfactual reward that explicitly optimizes the decision utility of forensic states, bridging the gap between representing visual evidence and actually using it for authenticity decisions.
- Extensive experiments on HydraFake demonstrate that COVE achieves state-of-the-art generalization across diverse evaluation scenarios, while comprehensive ablations and analyses validate the effectiveness of its key designs and the quality of its forensic reasoning.

2 Related Work

Generalizable deepfake detection. Image authenticity detectors exploit complementary visual signals, including blending artifacts (Shiohara & Yamasaki, 2022), local pixel relationships (Tan et al., 2024), and diffusion reconstruction errors (Wang et al., 2023). Pretrained representations support transfer across generators, as demonstrated by UniFD (Ojha et al., 2023), while Co-SPY (Cheng et al., 2025) combines semantic and features from individual pixels. Multimodal large language models extend detection to evidence description and reasoning: FakeVLM (Wen et al., 2025) learns artifact explanations, AIGI-Holmes (Zhou et al., 2025) combines visual experts with supervised and preference tuning, and FakeReasoning (Gao et al., 2026) couples CLIP/DINO feature fusion with forgery reasoning. X^2-DFD (Chen et al., 2025) constructs explanatory training data around the MLLM's existing capabilities and complementary specialist detectors. Veritas (Tan et al., 2026) learns structured forensic reasoning through pattern-aware reinforcement learning, and VIGIL (Li et al., 2026) plans inspection of individual regions and injects regional evidence into the reasoning

process. REFORM (Zhang et al., 2026) uses a reasoning curriculum with three stages and reinforcement learning for manipulation detection and grounding in paired images and text. Our work connects forensic representations learned from visual experts to training objectives that encourage their use in the verdict and preserve image dependence when learning from corrected reasoning trajectories.

Continuous and visually grounded reasoning. Coconut (Hao et al., 2025) feeds hidden states back as inputs to perform reasoning in continuous space, while CODI (Shen et al., 2025) distills explicit chains of thought into compact latent computation. For visual reasoning, LVR (Li et al., 2025) interleaves text with continuous visual states, Monet (Wang et al., 2025) learns visual latents through distillation and policy optimization over visual latents, and ILVR (Dong et al., 2026) selects perceptual targets conditioned on the evolving reasoning context. Most closely related to our representation learning, CoVT (Qin et al., 2025) supervises compact continuous visual tokens using complementary visual experts. COVE builds on this approach by aligning designated forensic token states with texture, semantic, and boundary representations, and uses a counterfactual token utility reward to encourage their contribution to the verdict. We further adapt the complete reasoning policy through correction and replay, retaining the aligned forensic states within each trajectory.

Counterfactual grounding and learning relative to a reference. CFPO (Yu et al., 2026) intervenes on visual values at salient cross-modal attention entries and optimizes policy divergence from the intervened branch. It motivates measuring visual dependence through controlled changes to a model's computation. COVE uses two complementary comparisons: CTUR measures the change in the likelihood of the correct label after forensic token input replacement, while image contrast compares a fixed teacher analysis under matched and donor images. The latter uses a logistic objective relative to a reference model, as in DPO (Rafailov et al., 2023). DPO compares responses for a fixed input. Our refinement objective compares image conditions for a fixed teacher trajectory and scores only its analysis and reflection spans. Subtracting the frozen base model's image gap targets the additional image selectivity learned during correction.

3 METHOD

3.1 OVERVIEW

COVE integrates continuous visual forensic states into structured reasoning for generalizable deepfake detection (Figure 1). Given an image x and prompt u, the model generates a reasoning chain $y = (y_1, \ldots, y_T)$ and an authenticity verdict $a \in \mathrm{real}, \mathrm{fake}$, with continuous forensic states preserving fine-grained visual evidence throughout reasoning. Training proceeds in three stages: (1) **Continuous visual state learning** aligns forensic states with visual experts to capture diverse forensic cues; (2) **Counterfactual reward optimization** uses rule-based reinforcement learning to promote their use in decision making; (3) **Contrastive Forensic Grounding** contrasts semantically matched real–generated pairs to strengthen visual grounding. At inference, a single model generates the complete reasoning chain and verdict.

3.2 CONTINUOUS FORENSIC STATES WITH EXPERT SUPERVISION

Deepfake evidence often manifests as fine-grained visual cues that are difficult to fully capture with discrete language. While structured reasoning helps organize forensic analysis (Tan et al., 2026), verbal descriptions may compress or omit visual details critical for identifying subtle manipulations. We therefore augment textual reasoning with *continuous forensic states*, which encode visual evidence directly in the model's hidden space and are grounded in complementary expert representations. Let $\mathcal{E}$ denote a set of visual experts that provide heterogeneous forensic supervision. For each expert $e \in \mathcal{E}$, we designate a set of positions in the reasoning sequence as continuous forensic states and collect their final-layer hidden representations as $H^e_\theta(x, u, y) \in \mathbb{R}^{M_e \times d_{\mathrm{lm}}}$, where M_e denotes the number of states and d_{lm} the language model hidden dimension. Conditioned on the input image and preceding context, these states form compact representations of complementary forensic evidence that can be directly attended to by subsequent reasoning and prediction tokens. Since expert targets may differ in dimensionality and structure, we use an expert-specific projection D_e to map the corresponding forensic states into each target space, $\widehat{F}^e = D_e(H^e_\theta)$, where $F^e(x)$ denotes the representation provided by expert e. We jointly ground all forensic states through a

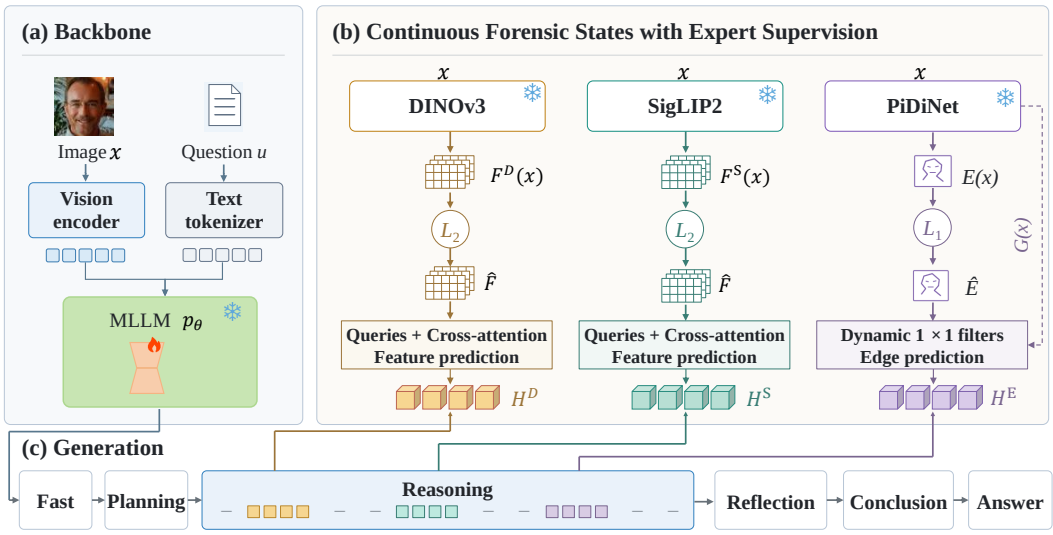


Figure 1: Continuous forensic states are learned under visual expert supervision and integrated into the LLM reasoning trajectory.

unified alignment objective:

$$\mathcal{L}_{\text{align}} = \sum_{e \in \mathcal{E}} \lambda_e \, \ell_e\Big(\widehat{F}^e, \text{sg}(F^e(x))\Big), \tag{1}$$

where ℓ_e measures the discrepancy in the corresponding target space, λ_e controls the contribution of each expert, and $\text{sg}(\cdot)$ denotes stop-gradient. The expert encoders remain frozen throughout training, so gradients from $\mathcal{L}_{\text{align}}$ update only the multimodal model and the associated projection modules. This unified formulation accommodates heterogeneous expert supervision while preserving complementary visual evidence in a shared continuous interface for subsequent forensic reasoning. Implementation details for expert-specific alignment are provided in Appendix A.1.

3.3 Counterfactual Reward for Forensic Utility

The alignment objective determines what the continuous forensic states encode, but does not ensure that the model actually uses them when making its final decision. We therefore complement representation alignment with a counterfactual objective that measures how much support for the correct verdict depends on these states, and optimize this signal through group-relative policy optimization.

Counterfactual Token Utility. Consider a reasoning trajectory o sampled under the factual inputs, and let $\mathcal{P}(o)$ denote its forensic-state positions. We construct a counterfactual counterpart by replacing the input embedding at each $p \in \mathcal{P}(o)$ with the vocabulary mean $\bar{e} = |\mathcal{W}|^{-1} \sum_{v \in \mathcal{W}} e_v$, while keeping the remaining prefix inputs and sequence structure fixed. We then recompute the affected hidden states and regenerate the continuation up to the answer boundary, allowing the intervention to propagate through subsequent reasoning (Figure 2a). Let c^F and c^{CF} denote the factual and counterfactual contexts immediately preceding the answer, respectively. We define the *Counterfactual Token Utility Reward* (CTUR) as:

$$R_{\text{CTUR}} = \log P_\theta(a^\star \mid x, u, c^F) - \log P_\theta(a^\star \mid x, u, c^{CF}), \tag{2}$$

where $a^\star$ is the ground-truth authenticity label and its log-likelihood is summed over all constituent tokens. A positive R_{CTUR} indicates that suppressing the forensic-state inputs reduces support for the correct verdict. Because the continuation is regenerated after intervention, CTUR captures not only the local contribution of the forensic states but also their propagated influence on subsequent reasoning.

Reward Composition. CTUR measures whether the verdict depends on the designated forensic states, but does not by itself enforce a correct or well-formed reasoning trajectory. We therefore

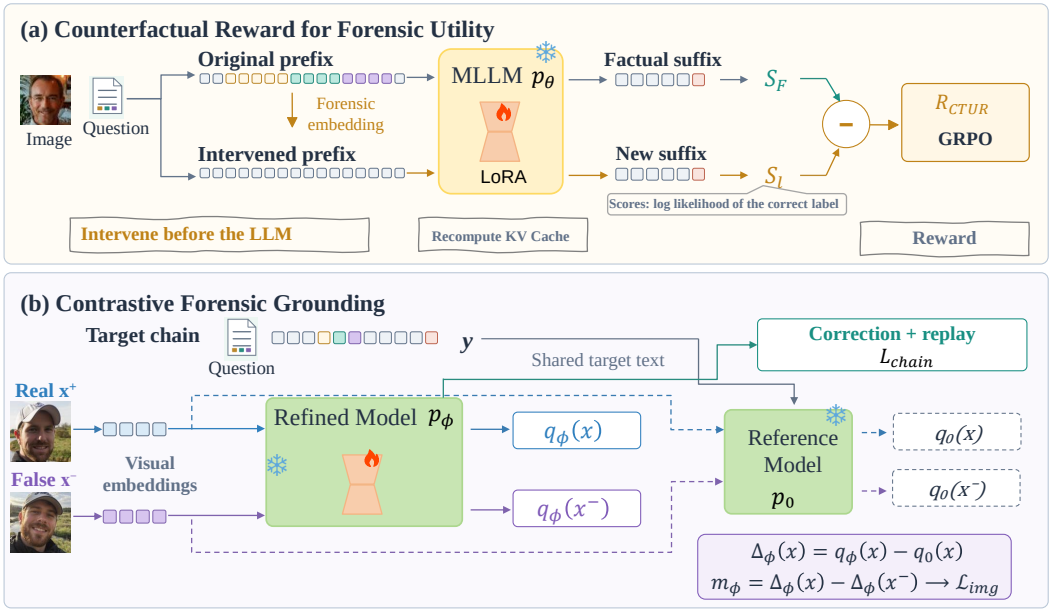


Figure 2: (a) CTUR measures forensic-state utility through counterfactual replacement and uses the resulting utility signal to optimize reasoning with GRPO. (b) Contrastive Forensic Grounding contrasts the same target reasoning under matched visual conditions to strengthen its dependence on authenticity-specific visual evidence.

combine it with the pattern and format rewards used for structured forensic reasoning (Tan et al., 2026):

$$R(o) = w_{\text{pat}} R_{\text{pat}}(o) + w_{\text{fmt}} R_{\text{fmt}}(o) + w_{\text{ctur}} R_{\text{CTUR}}(o), \tag{3}$$

where R_{pat} rewards answer correctness together with the desired reasoning behavior, R_{fmt} enforces the prescribed output structure, and w_{pat}, w_{fmt}, and w_{ctur} control their relative contributions. Thus, the former rewards provide task- and structure-level supervision, while CTUR explicitly rewards the utility of forensic states for the final decision.

Group-Relative Policy Optimization. We optimize the resulting reward with GRPO. For each image–prompt pair (x, u), we sample a group of G reasoning trajectories $\{o_i\}_{i=1}^{G}$ from the rollout policy $\pi_{\theta_{\text{old}}}$. Each trajectory receives the combined reward $R_i = R(o_i)$ from Eq. equation 3. We obtain its group-relative advantage by normalizing rewards within the sampled group $A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon}$. This normalization converts the heterogeneous reward components into a relative signal that favors trajectories exhibiting stronger forensic-state utility together with correct and structured reasoning. Let $r_i(\theta) = \frac{\pi_\theta(o_i|x,u)}{\pi_{\theta_{\text{old}}}(o_i|x,u)}$ denote the policy ratio. The policy is optimized with the clipped group-relative objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(x,u)} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Big(\min \left[r_i(\theta) A_i,\ \text{clip}\left(r_i(\theta), 1-\epsilon_l, 1+\epsilon_h\right) A_i \right] - \beta' D_{\text{KL}}\left(\pi_\theta(\cdot \mid x, u) \,\|\, \pi_{\text{ref}}(\cdot \mid x, u)\right) \Big) \Bigg], \tag{4}$$

where π_{ref} is the reference policy and β' controls the KL regularization.

3.4 Contrastive Forensic Grounding

Although counterfactual optimization promotes the use of forensic evidence, p_0 may still struggle with challenging cases. We therefore refine it with complete-chain supervision and contrastive visual

grounding, correcting its reasoning while strengthening its dependence on visual evidence. The refined model p_ϕ is initialized from p_0, which remains frozen as a reference.

Correction and Replay. We partition training examples according to whether the reasoning trajectories generated by p_0 lead to the correct verdict. For unsuccessful trajectories, we construct structured teacher trajectories with the ground-truth verdict to form a correction set $\mathcal{H}$; for successful trajectories, we retain the original generations as a replay set $\mathcal{B}$. The former guides the model toward more reliable forensic reasoning, while the latter preserves its acquired reasoning behavior. Both retain the complete reasoning structure, including the continuous forensic states introduced in Section 3.2.

Complete Chain Supervision. Given an image x, prompt u, and target chain y, we train p_ϕ autoregressively with teacher forcing. For each target position t, the token loss is

$$\ell_t = -\log p_\phi(y_t \mid x, u, y_{<t}). \tag{5}$$

Because a complete chain contains a substantially longer reasoning trace than its final verdict, we normalize their contributions separately. Let $\mathcal{A}$ denote the answer positions, $\mathcal{V}$ the forensic-state positions, and $\overline{\mathcal{A}}$ the remaining target positions. We define

$$\mathcal{L}_{\text{chain}} = \frac{1}{|\overline{\mathcal{A}}|}\sum_{t\in\overline{\mathcal{A}}}\ell_t + \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}\ell_t + \frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}}\ell_t. \tag{6}$$

The continuous forensic states remain embedded throughout the reasoning trajectory, allowing complete-chain supervision to jointly refine reasoning.

Contrastive Evidence Grounding. Chain supervision alone does not ensure that forensic reasoning is grounded in visual evidence. We therefore construct semantically matched image pairs (x^+, x^-) that preserve identity and high-level content while differing in authenticity. Specifically, given a real image x^+, an MLLM produces its semantic description d, which conditions a generative model to synthesize its counterpart x^-. This controlled contrast isolates authenticity-related visual differences while minimizing semantic confounds:

$$q_\phi(x, u, y) = \frac{1}{|\mathcal{E}(y)|}\sum_{t\in\mathcal{E}(y)}\log p_\phi(y_t \mid x, u, y_{<t}). \tag{7}$$

which measures the support for the same forensic reasoning under visual condition x. We analogously compute q_0 using the frozen reference model p_0.

Contrastive Margin Loss. Simply encouraging $q_\phi(x^+, u, y) > q_\phi(x^-, u, y)$ may reinforce a preference already present in p_0, without ensuring that refinement itself strengthens visual grounding. We instead compare how much support for the same reasoning changes relative to the frozen base model. Define the refinement gain under image x as $\Delta_\phi(x) = q_\phi(x, u, y) - q_0(x, u, y)$. We then define the visual grounding margin as the difference between the gains under the matched image pair $m_\phi = \Delta_\phi(x^+) - \Delta_\phi(x^-)$. The corresponding margin loss is

$$\mathcal{L}_{\text{img}} = \log(1 + \exp(-\beta m_\phi)), \tag{8}$$

where β controls the sharpness of the comparison. If refinement improves the likelihood of the teacher reasoning equally under both images, then $m_\phi = 0$, indicating no additional preference for its supporting visual evidence. In contrast, $m_\phi > 0$ indicates a larger grounding gain under the matched real image. The objective therefore encourages forensic reasoning to depend on its supporting visual evidence rather than merely increasing the likelihood of the teacher chain.

Joint Refinement. The two objectives play complementary roles: chain supervision teaches *what* reasoning should be corrected, while contrastive grounding encourages the correction to remain tied to *which visual evidence* supports it. For a correction batch $B_\mathcal{H}$ and a replay batch $B_\mathcal{B}$, we optimize

$$\mathcal{L}_{\text{refine}} = \frac{1}{|B_\mathcal{H}|}\sum_{i\in B_\mathcal{H}}(\mathcal{L}_{\text{chain},i} + \lambda_{\text{img}}\mathcal{L}_{\text{img},i}) + \frac{\lambda_{\text{rep}}}{|B_\mathcal{B}|}\sum_{j\in B_\mathcal{B}}\mathcal{L}_{\text{chain},j}, \tag{9}$$

where λimg and λ_{rep} control contrastive grounding and replay, respectively. The reference model p_0 remains frozen, while p_ϕ is optimized under both visual conditions. At inference, only p_ϕ is required to generate the complete reasoning trajectory and verdict.

Table 1: Performance comparison (Acc., %) on HydraFake. In-domain (ID) results are averaged over five subsets. Avg. excludes ID for methods without ID evaluation. COVE-MINI is trained only on FF++, StyleGAN, Stable Diffusion XL, and FFHQ. Excluding the cold-start model, the best results are **bolded** and the second best are underlined.

Method	ID	Cross-Model						Cross-Forgery						Cross-Domain						Avg.
		ADF	FLUX	StarryAI	MAGI-1	HART	Infinity	StGANv2	ICLight	CodeF.	Infinite Y.	PuLID	FaceAda.	DeepFace.	Infinite Y.	Dreamina	HailuoAI	GPT-4o	FFIW	
Small vision models																				
F3Net (ECCV'20)	85.3	86.7	87.8	78.6	85.0	86.0	82.9	41.3	48.9	71.9	84.9	85.5	72.6	57.7	78.5	55.6	68.6	66.2	66.4	73.2
UniFD (CVPR'23)	82.7	90.7	93.8	82.5	73.0	94.4	90.7	61.8	81.9	75.4	73.7	68.1	81.3	67.4	67.3	80.5	75.2	73.3	67.5	78.0
IID (CVPR'23)	83.4	83.3	82.8	80.0	80.2	81.1	82.2	41.4	53.3	79.7	81.8	81.8	73.7	65.2	69.9	63.8	63.3	63.8	64.2	72.4
FreqNet (AAAI'24)	66.8	60.3	76.7	59.0	69.2	77.1	75.1	33.1	73.1	70.3	72.8	77.4	67.7	50.6	67.0	62.1	59.3	58.3	51.2	64.6
ProDet (NeurIPS'24)	90.5	92.6	94.2	88.2	91.9	93.8	93.1	56.3	58.6	80.8	88.1	91.0	83.3	58.1	82.9	71.3	75.6	66.3	74.1	80.6
NPR (CVPR'24)	75.6	68.8	91.2	59.5	82.6	91.3	84.0	47.7	67.8	60.6	79.8	89.0	67.7	52.6	73.0	76.6	62.3	50.2	46.0	69.8
AIDE (ICLR'25)	80.4	68.8	86.3	64.0	88.9	95.4	76.0	56.7	79.2	86.1	74.2	62.4	75.7	59.7	67.9	49.7	58.0	51.9	59.2	70.6
Co-SPY (CVPR'25)	86.3	93.5	95.5	85.3	93.3	96.6	95.3	77.0	92.5	88.6	90.6	79.1	87.3	67.6	80.0	82.5	74.0	79.5	64.3	84.7
D^3 (CVPR'25)	87.3	93.6	95.6	91.3	90.7	95.8	95.5	62.4	71.6	82.9	80.0	82.4	73.7	69.7	74.6	78.1	70.9	80.8	64.3	81.1
Effort (ICML'25)	94.7	82.8	96.5	78.0	90.5	97.8	98.3	64.7	**94.8**	89.7	89.5	92.9	88.0	64.8	82.2	61.5	66.4	53.8	74.0	82.2
Generic MLLMs																				
Qwen2.5-VL-7B	51.2	50.0	50.0	49.7	50.0	52.0	52.9	50.5	56.7	50.7	53.6	54.5	51.6	50.7	53.6	80.2	67.5	52.5	50.5	54.1
InternVL3-8B	54.0	54.0	49.8	49.0	56.6	55.8	57.2	62.9	54.2	62.9	63.6	54.8	67.7	54.4	67.1	77.1	66.5	47.4	51.8	58.3
MiMo-VL-7B	63.8	74.5	77.1	82.5	60.3	82.4	81.4	48.7	82.6	76.4	79.7	78.4	82.8	57.7	75.6	79.4	70.7	67.7	54.9	72.5
GLM-4.1V-9BThink	56.4	55.2	52.3	50.5	51.6	68.4	60.7	54.3	68.4	63.3	65.7	55.1	81.0	58.7	72.7	83.7	69.2	52.0	53.9	61.7
GPT-4o	53.5	57.7	52.0	51.4	59.9	81.2	54.8	66.4	58.9	52.5	64.4	60.9	55.5	49.4	62.0	90.7	73.7	58.0	52.8	60.8
Gemini-2.5-Pro	72.2	64.9	92.4	82.8	62.5	93.4	93.2	73.7	83.3	87.4	85.5	84.7	85.6	67.2	75.6	87.5	82.4	70.9	53.0	78.9
Forgery detectors based on MLLMs																				
M2F2-Det (CVPR'25)	-	56.0	57.7	59.8	61.8	61.3	55.4	78.9	65.5	80.0	57.4	57.5	76.3	73.0	56.3	67.2	50.6	53.0	70.6	63.2
FakeShield (ICLR'25)	-	64.3	64.0	61.5	63.1	61.8	63.3	64.0	57.3	60.9	58.1	63.6	63.7	50.2	83.8	53.8	51.3	53.9	55.6	60.8
SIDA-7B (CVPR'25)	-	97.3	97.7	79.5	59.3	98.5	95.0	59.8	60.6	62.3	89.7	94.4	63.3	50.4	81.9	80.0	78.0	68.9	57.3	76.3
SIDA-13B (CVPR'25)	-	80.7	78.5	54.8	52.5	91.3	82.4	63.7	61.2	68.2	56.7	67.1	84.3	60.8	58.2	88.3	74.0	74.1	59.9	69.8
FFAA (arXiv'24)	-	55.1	50.9	72.9	63.5	60.8	57.6	82.7	70.9	71.8	58.4	62.4	86.0	67.7	58.4	55.3	59.2	49.6	68.3	64.0
FakeVLM (NeurIPS'25)	-	78.2	78.5	77.0	74.5	76.5	76.8	70.8	76.2	76.2	76.9	76.5	77.7	**75.7**	83.6	81.5	80.8	78.7	74.5	77.3
Veritas (ICLR'26)	97.3	94.8	**99.8**	97.0	**99.9**	**99.9**	**99.9**	90.3	75.7	97.0	91.8	95.1	91.7	58.6	84.1	92.3	90.2	**89.2**	78.5	90.7
VIGIL (arXiv'26)	**98.2**	96.3	99.5	96.2	97.2	99.1	99.6	95.5	88.0	**97.1**	94.9	95.9	93.4	67.5	86.9	94.1	91.5	85.7	73.5	92.1
COVE-MINI	-	90.5	99.3	99.2	73.1	95.2	96.8	83.9	91.4	90.1	88.7	90.4	88.0	70.2	76.6	**94.9**	**92.9**	86.0	71.8	87.7
COVE (cold-start)	97.1	88.0	99.7	96.8	99.7	99.8	99.9	87.3	70.9	95.1	86.3	93.6	86.0	58.3	74.3	95.9	94.3	83.9	77.1	88.6
COVE (base stage)	97.2	96.0	99.7	99.3	99.8	99.7	99.8	97.3	92.7	96.0	97.3	96.9	**97.7**	60.2	87.9	93.7	90.6	86.8	80.0	93.1
COVE (ours)	97.5	**98.5**	**99.8**	**99.5**	99.6	**99.9**	99.8	**98.4**	93.2	**97.1**	**98.7**	**97.9**	97.0	69.4	**88.8**	93.8	90.7	87.0	**80.2**	**94.0**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks and data. HydraFake (Tan et al., 2026) contains five in-domain (ID) subsets and six subsets each for cross-model (CM), cross-forgery (CF), and cross-domain (CD) evaluation. These groups test changes in generator, manipulation category, and real/fake image source, respectively. Correction and replay use reasoning trajectories from the training split. We evaluate the model on the HydraFake test sets.

Metrics. Following established protocols in deepfake detection (Tan et al., 2026), we adopt accuracy (Acc.) as the primary metric. We additionally report precision and recall in Appendix B.

Implementation details. We implement COVE with InternVL3-8B (Zhu et al., 2025). Base training uses LoRA (Hu et al., 2022) with rank 128 and scale 256. SFT and direct preference optimization (DPO) (Rafailov et al., 2023) run for three and two epochs, respectively, with a learning rate of 5×10^{-5} and a batch size of 64. The experts remain frozen, and the alignment weights are $\alpha_1 = 2$ and $\alpha_2 = 1$. We set $\lambda_e = 100$ for $e \in \mathcal{R}$ and $\lambda_{\text{edge}} = 0.025$. GRPO runs for two epochs with a learning rate of 10^{-6} and a batch size of 16, with $G = 4$ rollouts per image, temperature 1.0, and $\beta' = 0$. The reward weights are $w_{\text{pat}} = 1$, $w_{\text{fmt}} = 0.025$, and $w_{\text{ctur}} = 0.3$. Each stage starts from the preceding checkpoint. The correction stage uses LoRA with rank 32 and scale 64. We train with AdamW for 128 steps at a learning rate of 5×10^{-5}, using 16 correction and 24 replay examples per step. We set $\lambda_{\text{img}} = 1$, $\lambda_{\text{rep}} = 0.5$, and $\beta = 10$.

4.2 MAIN RESULTS

Benchmark performance. Table 1 reports 94.0% Overall accuracy for COVE, with 97.5/99.5/97.1/85.0 across ID/CM/CF/CD. The complete model exceeds VIGIL by 1.9 percentage points and Veritas with the same backbone by 3.3%. Relative to the earlier base result of 93.1%, the

overall improvement is 0.9 points. All four group averages improve, with the largest increase on CD from 83.2% to 85.0%. These results support the combination of continuous forensic representations, decision supervision, and reasoning refinement across different sources of distribution shift.

Performance across subsets. DeepFaceLab improves from 60.2% to 69.4% accuracy, while CD-InfiniteYou increases from 87.9% to 88.8%. The gain is smaller on FFIW, from 80.0% to 80.2%, and FaceAdapter decreases from 97.7% to 97.0%. The benefit therefore varies across manipulation types and image sources. On DeepFaceLab, fake recall of 99.1% accompanies precision of 62.2%, indicating frequent false alarms despite detecting most fake images. These differences motivate examining precision and recall alongside accuracy, and measuring both corrected and regressed predictions in the controlled comparisons.

4.3 Ablation Studies

We examine forensic experts, the CTUR reward, reasoning refinement, and token count. The expert and reward ablations evaluate the base training stages before refinement.

Table 2: Ablation of forensic expert tokens. D, S, and P denote DINOv3, SigLIP2, and PiDiNet. Check marks indicate enabled groups.

D	S	P	ID	CM	CF	CD
			97.3	98.6	90.3	82.2
✓			97.3	98.7	92.4	82.4
	✓		97.2	98.6	91.9	82.3
		✓	97.3	98.6	92.0	82.4
✓	✓		97.3	98.8	94.8	82.6
✓		✓	97.2	99.0	95.3	82.9
	✓	✓	97.3	98.9	94.9	82.8
✓	✓	✓	97.2	99.1	96.3	83.2
Δ forensic			-0.1	+0.5	+6.0	+1.0

Effect of forensic experts. Table 2 compares DINOv3, SigLIP2, and PiDiNet individually and in combination. CF accuracy increases from 90.3% without expert tokens to 96.3% with all three groups. Each expert improves CF accuracy on its own, and the complete combination exceeds every individual expert and pair on CF. The complete combination also gives the highest CD accuracy, at 83.2%, compared with 82.2% without experts and 82.9% for the strongest pair. These results support the complementary value of the three forensic experts.

Effect of the CTUR reward. Table 3 evaluates whether explicitly optimizing the decision utility of forensic states improves detection. The two GRPO variants share the same initialization and training settings, differing only in the inclusion of CTUR. Adding CTUR improves overall accuracy from 91.2% to 93.1%, with particularly clear gains of 3.1% and 2.5% on CF and CD, respectively. This improvement indicates that learning informative forensic representations alone is insufficient; explicitly encouraging the model to rely on these states provides additional benefits under distribution shifts.

Table 3: Training stages and CTUR (%). Both RL variants start from the same DPO checkpoint and differ only in the CTUR reward.

Endpoint	ID	CM	CF	CD	Overall
SFT	96.4	97.6	82.5	77.1	86.3
Continued SFT	96.6	97.8	83.4	77.9	86.9
DPO	97.1	97.3	86.5	80.6	88.6
RL w/o CTUR	97.1	98.7	93.2	80.7	91.2
RL with CTUR	97.2	99.1	96.3	83.2	93.1

Table 4: Reasoning refinement. Corr. and Reg. count corrections and regressions from the same base. Invalid outputs count as incorrect.

Variant	Overall	CD	Corr.	Reg.
Frozen base p_0	93.1	83.2	0	0
w/o image contrast	93.5	83.9	644	420
w/o replay	93.6	85.3	1048	768
Raw gap	93.7	84.4	753	411
Answer CE	93.4	83.7	611	445
COVE (ours)	94.0	85.0	968	403

Effect of Contrastive Forensic Grounding. Table 4 validates the effectiveness of contrastive forensic grounding. Compared with chain supervision alone (w/o image contrast), our full objective improves Overall/CD accuracy from 93.5/83.9% to 94.0/85.0%, showing the benefit of grounding refinement in visual evidence. Using the raw image gap reduces performance to 93.7/84.4%, supporting the reference-relative formulation, while answer-only supervision further drops to 93.4/83.7%, highlighting the importance of refining the complete reasoning trajectory. Replay additionally reduces regressions from 768 to 403 while retaining substantial corrections. Together, these results support the key designs of contrastive forensic grounding for effective and stable reasoning refinement.

Number of visual tokens. We vary the number of tokens per expert to examine how the capacity of continuous forensic states trades off accuracy and computation (Figure 3). Increasing K from 2 to 4 substantially improves accuracy from 91.9% to 94.0%, suggesting that a small number of states may be insufficient to capture complementary fine-grained forensic evidence. Beyond this point, however, the benefit quickly saturates: using six tokens yields only a marginal 0.1-point improvement, while the time overhead increases from 29% to 37%. Further increasing K to 8 raises the overhead to 43% and slightly reduces accuracy to 93.7%, indicating that additional states do not necessarily provide more useful forensic information.

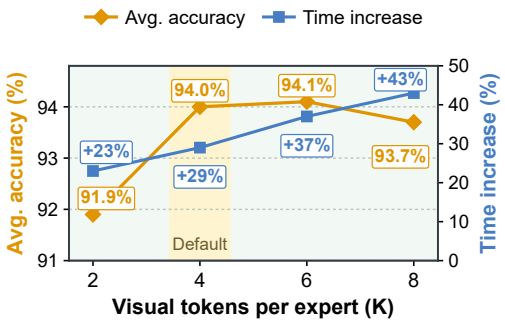


Figure 3: Accuracy and time versus tokens per expert. Time increase is relative to $K = 0$.

Table 5: Evaluation of reasoning quality.

Model	Score Evaluation		ELO Rating
	GPT-4o	Gemini-2.5-Pro	
MiMo-VL-7B	3.0731	2.5785	531.6
GPT-4o	2.4718	2.1619	652.3
Gemini-2.5-Pro	4.1681	4.0070	861.7
Veritas	4.6479	4.4214	1033.2
COVE (cold-start)	4.2169	4.1374	1152.4
COVE (base stage)	4.8471	4.6158	1376.5
COVE (ours)	**4.8925**	**4.6502**	**1392.3**

Table 6: Robustness on compression and blur.

Method	Orig.	JPEG (QF)			Blur (σ)	
		90	70	50	1.0	2.0
UniFD	78.0	77.1	75.4	72.2	76.5	75.4
IID	72.4	71.4	70.1	69.7	72.1	71.9
ProDet	80.6	80.1	79.4	78.7	79.7	77.8
Co-SPY	84.7	84.5	83.4	81.0	82.7	77.0
Effort	82.2	73.1	79.6	66.3	79.1	70.4
Veritas	90.7	90.1	88.7	87.4	88.8	84.3
COVE	**94.0**	**93.2**	**92.8**	**92.6**	**92.9**	**89.5**

Evaluation of reasoning quality. We evaluate reasoning quality using GPT-4o and Gemini-2.5-Pro as judges and Elo ratings from pairwise comparisons. As shown in Table 5, COVE achieves the highest scores across all three metrics. Refinement improves the two judge scores from 4.8471/4.6158 to 4.8925/4.6502 and Elo from 1376.5 to 1392.3, demonstrating consistent improvements in reasoning quality.

Robustness to compression and blur. We evaluate robustness to JPEG compression and Gaussian blur, which degrade fine-grained forensic cues. As shown in Table 6, COVE maintains 92.6% accuracy at JPEG quality 50 and 89.5% under blur $\sigma = 2.0$, while consistently outperforming all compared methods across degradation levels. These results demonstrate strong robustness to image degradation, with severe blur posing a greater challenge than compression.

5 Conclusion

We presented COVE, a unified framework that connects continuous forensic representations, decision utility, and reasoning refinement. Expert supervision grounds forensic states in complementary visual cues, while CTUR encourages their contribution to the verdict. Refinement combines complete corrected chains, image comparison relative to a frozen base model, and replay of existing correct responses. On HydraFake, COVE achieves 94.0% overall accuracy and 85.0% cross-domain accuracy. Ablations show that image comparison and supervision of intermediate reasoning improve detection, while replay reduces regressions. These results support treating forensic representation, decision use, and learning from errors as complementary parts of generalizable deepfake detection.

Limitation and future direction. COVE relies on fine-grained visual evidence and thus remains sensitive to severe image degradation; its contrastive grounding also depends on the quality of semantically matched visual conditions. Improving robustness to visual corruption, constructing more diverse and faithful contrastive pairs, and extending evidence-grounded reasoning to broader manipulation types and real-world scenarios are promising directions for future research.

AI USE STATEMENT

Generative AI models were used in the COVE pipeline to construct semantically matched visual conditions for contrastive forensic grounding. Specifically, an MLLM was used to derive semantic descriptions from real images, which were then used to guide image generation while preserving high-level semantic content and varying authenticity. AI models were also used as judges for reasoning-quality evaluation. These AI-assisted components are used for training-data construction and evaluation rather than as sources of ground-truth authenticity labels. AI tools additionally provided language assistance during manuscript preparation. The authors take responsibility for the final manuscript, experimental results, and accompanying artifacts.

ETHICS STATEMENT

This work aims to advance generalizable deepfake detection and evidence-grounded forensic reasoning. COVE is developed and evaluated on existing deepfake research data and uses generative models to construct controlled visual conditions for studying and improving the grounding of forensic reasoning. These generated samples are used solely for research on manipulation detection and reasoning, rather than for impersonation, fraud, or other malicious applications. Given the dual-use nature of deepfake research, the proposed method focuses on detecting and analyzing manipulated content and does not provide tools or pipelines intended to facilitate deceptive media generation. We follow the access and usage terms of the datasets and models used in our experiments. Additional details on datasets and experimental protocols are provided in the Appendix.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the COVE architecture, training objectives, data construction procedure, and evaluation protocols in the main paper and Appendix. This includes the construction of continuous forensic states, expert supervision, Counterfactual Token Utility Reward (CTUR), Contrastive Forensic Grounding, training configurations, and evaluation settings. We also report component-wise ablations and controlled interventions to facilitate verification of the proposed designs. The supplementary material contains additional implementation details and experimental analyses necessary for reproducing the reported results.

REFERENCES

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X^2-DFD: A Framework for eXplainable and eXtendable Deepfake Detection. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.06126.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025. URL https://arxiv.org/abs/2503.18286.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved Latent Visual Reasoning with Selective Perceptual Modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1351. URL https://aclanthology.org/2026.acl-long.1351/.

Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Muxi Diao, Lei Chen, Kongming Liang, and Zhanyu Ma. Toward Generalizable Forgery Detection and Reasoning. *IEEE Transactions on Image Processing*, 35:3395–3410, 2026. doi: 10.1109/TIP.2026.3673967. URL https://doi.org/10.1109/TIP.2026.3673967.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 105–116, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2412.06769.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, Wenming Yang, and Jiaya Jia. FFAA: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*, 2024.

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

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent Visual Reasoning. *arXiv preprint arXiv:2509.24251*, 2025. URL https://arxiv.org/abs/2509.24251.

Xinghan Li, Junhao Xu, and Jingjing Chen. VIGIL: Part-Grounded Structured Reasoning for Generalizable Deepfake Detection. *arXiv preprint arXiv:2603.21526*, 2026. URL https://arxiv.org/abs/2603.21526.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. Chain-of-Visual-Thought: Teaching VLMs to See and Think Better with Continuous Visual Tokens. *arXiv preprint arXiv:2511.19418*, 2025. URL https://arxiv.org/abs/2511.19418.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing Chain-of-Thought into Continuous Space via Self-Distillation. *arXiv preprint arXiv:2502.21074*, 2025. URL https://arxiv.org/abs/2502.21074.

Kaede Shiohara and Toshihiko Yamasaki. Detecting Deepfakes with Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18699–18708, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikainen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5117–5127, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Su_Pixel_Difference_Networks_for_Efficient_Edge_Detection_ICCV_2021_paper.html.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable Deepfake Detection via Pattern-Aware Reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.21048.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in Latent Visual Space Beyond Images and Language. *arXiv preprint arXiv:2511.21395*, 2025. URL https://arxiv.org/abs/2511.21395.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the Fake: Large Multimodal Model-Based Synthetic Image Detection with Artifact Explanation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.14905.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: Counterfactual Policy Optimization for Multimodal Reasoning. *arXiv preprint arXiv:2606.23206*, 2026. URL https://arxiv.org/abs/2606.23206.

Yuchen Zhang, Yaxiong Wang, Kecheng Han, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. Cultivating Forensic Reasoning for Generalizable Multimodal Manipulation Detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28545–28560. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1316. URL https://aclanthology.org/2026.acl-long.1316/.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025. URL https://arxiv.org/abs/2507.02664.

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

A Additional method details

A.1 Expert Supervision Alignment

We instantiate the expert set $\mathcal{E}$ in Section 3.2 with three complementary visual representations: DINOv3 (Siméoni et al., 2025) for local appearance, SigLIP2 (Tschannen et al., 2025) for semantic structure, and PiDiNet (Su et al., 2021) for boundary information. These experts provide different forms of supervision: DINOv3 and SigLIP2 produce sequences of visual features, whereas PiDiNet provides spatial features and an explicit boundary target. We accordingly instantiate the projection D_e and alignment loss ℓ_e in Eq. equation 1 according to the structure of each target space. All expert encoders are kept frozen during training.

Representation targets. Let $\mathcal{R} = \{\mathrm{DINO}, \mathrm{SigLIP}\} \subset \mathcal{E}$ denote the representation-based experts. Given an input image x, expert $e \in \mathcal{R}$ produces a target feature sequence

$$F^e(x) \in \mathbb{R}^{N_e \times d_e}, \tag{10}$$

where N_e and d_e denote the sequence length and feature dimension of the corresponding expert representation. Since these target spaces differ from the language-model hidden space in both length and dimensionality, we instantiate D_e as a lightweight query-based decoder. Specifically, a set of learned queries attends to the continuous forensic states H^e_θ through cross-attention and predicts a feature sequence in the corresponding expert space,

$$\widehat{F}^e = D_e(H^e_\theta) \in \mathbb{R}^{N_e \times d_e}. \tag{11}$$

This decoder allows the compact forensic states to reconstruct a variable-length expert representation without requiring a one-to-one correspondence between continuous states and expert features.

We align the predicted and target representations after row-wise ℓ_2 normalization. Denoting row normalization by $\mathrm{Norm}(\cdot)$, the representation alignment objective is

$$\mathcal{L}_{\mathrm{rep}} = \sum_{e \in \mathcal{R}} \lambda_e \left\| \mathrm{Norm}\left(\widehat{F}^e\right) - \mathrm{Norm}(\mathrm{sg}(F^e(x))) \right\|_F^2. \tag{12}$$

Normalization emphasizes agreement in feature direction while reducing sensitivity to differences in feature magnitude across expert spaces.

Boundary target. Boundary evidence has a different structure from sequence-based representations: it is spatially localized and is naturally expressed as a dense prediction over the image. We therefore instantiate the corresponding projection using dynamic filters conditioned on the boundary forensic states. Let $G(x) \in \mathbb{R}^{H \times W \times d_g}$ denote an intermediate spatial feature map extracted by the frozen boundary expert. Each boundary state $h^{\mathrm{edge}}_k \in H^{\mathrm{edge}}_\theta$ is mapped to the parameters of a dynamic 1×1 filter,

$$w_k = \phi_{\mathrm{edge}}\left(h^{\mathrm{edge}}_k\right), \tag{13}$$

where ϕ_{edge} is a learned projection. Applying these filters to $G(x)$ yields state-conditioned boundary responses, which are aggregated into the predicted boundary map

$$\widehat{E} = \sigma\left(\frac{1}{M_{\mathrm{edge}}} \sum_{k=1}^{M_{\mathrm{edge}}} w_k * G(x)\right), \tag{14}$$

where M_{edge} is the number of boundary forensic states, $*$ denotes convolution, and $\sigma(\cdot)$ is the sigmoid function. We supervise the prediction with the frozen boundary target $E(x)$ using

$$\mathcal{L}_{\mathrm{edge}} = \left\| \widehat{E} - \mathrm{sg}(E(x)) \right\|_1. \tag{15}$$

Unlike the representation targets, which encourage agreement in a learned feature space, this objective directly grounds the corresponding forensic states in spatially localized boundary evidence.

Overall expert alignment. Combining the representation and boundary objectives gives the concrete instantiation of the unified alignment objective in Eq. equation 1:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{rep}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}. \tag{16}$$

The stop-gradient operator prevents the alignment objective from updating the expert branches. Consequently, the expert representations provide stationary visual targets, while gradients from $\mathcal{L}_{\text{align}}$ optimize the multimodal model and the associated projection modules. Despite their different native structures, all expert signals therefore serve the same role: grounding compact continuous forensic states in complementary visual evidence that remains directly accessible to subsequent reasoning.

A.2 Training the base policy

Let $p_\theta(y \mid x, u)$ denote the likelihood of a complete trajectory. The supervised objective is

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,u,y)\sim\mathcal{D}_1} \sum_t \log p_\theta(y_t \mid x, u, y_{<t}). \tag{17}$$

For preference triples $(x, u, y_w, y_l) \sim \mathcal{D}_2$, preferred trajectories have better reasoning or a correct verdict relative to rejected trajectories. Using the SFT policy p_{SFT} as reference, DPO minimizes

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}_2} \log \sigma\left(\beta_D \left[\log \frac{p_\theta(y_w \mid x, u)}{p_{\text{SFT}}(y_w \mid x, u)} - \log \frac{p_\theta(y_l \mid x, u)}{p_{\text{SFT}}(y_l \mid x, u)}\right]\right). \tag{18}$$

Both stages include $\mathcal{L}_{\text{align}}$, with weights α_1 and α_2, respectively. The decoders are trained alongside the policy during alignment, and the visual teachers remain frozen.

For GRPO, sample G complete rollouts for each image and its prompt. Let p_{old} denote the policy used to sample this group. With rollout rewards R_i, define

$$A_i = \frac{R_i - \text{mean}_j\, R_j}{\text{std}_j\, R_j + \epsilon_A}, \qquad \rho_{i,t} = \frac{p_\theta(y_{i,t} \mid x, u, y_{i,<t})}{p_{\text{old}}(y_{i,t} \mid x, u, y_{i,<t})}. \tag{19}$$

The clipped policy objective, written as a loss, is

$$\begin{aligned} \mathcal{L}_{\text{GRPO}} &= -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_t s_{i,t}\right], \\ s_{i,t} &= \min(\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon_c, 1+\epsilon_c)A_i)\,. \end{aligned} \tag{20}$$

The default GRPO setting uses $\beta' = 0$. The reward combines the structured pattern and format terms with CTUR. The alignment loss is disabled in this stage. Counterfactual intervention and reward evaluation are detached from backpropagation: their effect enters through the rollout advantages. The prefix KV cache is recomputed after input replacement, and complete label likelihoods are used in both branches.

A.3 Correction and replay data

Each refinement run starts from the same base policy p_0. We partition training examples according to whether the reasoning trajectories generated by p_0 lead to the correct verdict. For unsuccessful trajectories, we construct a fixed complete teacher response whose final verdict matches the ground-truth label, forming the correction set. For successful trajectories, we retain the base model's original complete response for replay. Training images on which the base policy is correct but an earlier refinement checkpoint becomes incorrect can additionally be included in the replay set with their original base-policy trajectories. Unparseable predictions are treated as incorrect for this selection. All correction and replay examples are drawn from the training split.

For contrastive forensic grounding, we construct a semantically matched visual counterpart for each correction image. Given a real image x^+, an MLLM first produces a semantic description d capturing its identity and high-level visual content. We then condition a generative model on d to synthesize a corresponding image x^- that preserves the high-level semantics of x^+ while differing in authenticity. This produces the paired visual conditions (x^+, x^-) used by the contrastive grounding objective.

For each pair, the target reasoning trajectory y and task prompt u are held fixed, so that the two branches differ only in their visual condition. Dynamic preprocessing is controlled to maintain compatible image-token layouts across the pair. The implementation verifies matching prompt token IDs and embedding shapes, while the image embeddings differ between x^+ and x^-. During paired evaluation, both branches are conditioned on the same preceding target text at every scored position. This allows the contrastive objective to measure how strongly the same forensic reasoning is supported under the matched and counterfactual visual conditions.

A.4 TOKEN SELECTION AND PAIRED COMPUTATION

A mask selects the token positions included in a loss or score. The answer mask selects the final real/fake completion after any shared token prefix. All other target tokens, including structural delimiters and the stopping token, belong to the mask for the rest of the chain. The visual mask includes every occurrence of the twelve registered forensic token IDs, and overlaps with the mask for the rest of the chain. Each of the three CE terms in Equation 6 is averaged over its own mask before summation.

The evidence mask selects the analysis positions $\mathcal{E}(y)$ used in Equation 7. It maps the corresponding text spans to token positions using offsets that map characters to tokens. For each reasoning or reflection span, it begins after the first `<vis_pidinet_3>` token when present, and otherwise at the span's content start. It ends at the span's content boundary. Fast, planning, conclusion, and answer spans are excluded. The same masks are reused for both image conditions and both model states.

Evaluating the target chain under the original image supplies both its CE loss and analysis score. Evaluating it under the comparison image supplies the second score. Scores from the frozen reference p_0 are precomputed and stored without gradients. Both refined branches contribute gradients through the shared policy p_ϕ, while replay contributes chain supervision under its original image. Training settings are reported in Section 4.1.

B DETAILED DETECTION RESULTS

Tables 7–10 report accuracy, precision, and recall for every subset in the four evaluation groups, with fake as the positive class. Metrics use parsed outputs. Each group averages its constituent subsets. Overall accuracy averages the ID group score and the accuracies of the eighteen OOD subsets with equal weight. Comparisons without ID average the shared OOD subsets.

In-domain detection. Table 7 shows that COVE achieves the highest average accuracy of 97.5%, compared with 97.3% for Veritas and 94.7% for Effort. Accuracy reaches at least 99.5% on Hallo2, StyleGAN, and Midjourney, while FaceForensics++ and Facevid2vid remain harder at 92.4% and 95.9%. The latter two improve over the base stage by 0.7 and 0.8 percentage points. On Facevid2vid, 99.9% recall and 92.5% precision indicate that the remaining errors primarily arise from classifying real images as fake. Familiar forgery types therefore still require careful discrimination of genuine visual variation from artifacts.

Cross-model generalization. COVE reaches 99.5% average accuracy across the six unseen generators in Table 8, with every subset above 98%. MAGI-1, HART, and Infinity reach 99.6%, 99.9%, and 99.8%, respectively. The largest gain over the base stage occurs on Adobe Firefly (ADF), where accuracy rises from 96.0% to 98.5%. Recall increases from 92.7% to 99.3%, with precision decreasing from 99.3% to 97.7%. The improvement on this generator thus comes mainly from recovering previously missed synthetic images. The consistently high scores across generators support transfer of the learned forensic cues beyond the models represented in training.

Cross-forgery generalization. Table 9 evaluates transfer to manipulation types excluded from training. COVE achieves 97.1% average accuracy, compared with 96.3% for the base stage and 90.3% for Veritas. Five of the six subsets improve over the base stage, while FaceAdapter decreases from 97.7% to 97.0%. Relighting remains the most challenging case. On IC-Light, precision is 99.6% but recall is 86.7%, indicating that errors mainly reflect missed edits. Effort performs better on this subset, with 94.8% accuracy and 93.2% recall, compared with our 93.2% accuracy. Thus, strong transfer to unseen generators does not fully resolve sensitivity to changes in illumination that preserve facial identity and structure.

Cross-domain generalization. Changes in image sources remain the largest challenge (Table 10). COVE achieves the highest average accuracy of 85.0%, exceeding the base stage by 1.8 percentage points, with 93.8% on Dreamina and 90.7% on Hailuo AI. DeepFaceLab improves most, from 60.2% to 69.4%, but remains the weakest subset. Its precision increases from 55.7% to 62.2%, while recall stays high at 99.1%. The gain therefore primarily reflects fewer false alarms on real images. This error pattern differs from the missed edits on IC-Light. Generalization requires both recognizing subtle manipulations and accepting genuine images whose appearance differs from the training sources.

Table 7: Detailed in-domain results on HydraFake (%). Acc., P., and R. denote accuracy, precision, and recall, with fake as the positive class. Avg. is mean subset accuracy. The best averages are bold and the second best are underlined.

Method	FaceForensics++			Facevid2vid			Hallo2			StyleGAN			Midjourney			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net (ECCV'20)	84.5	80.5	90.9	89.6	83.6	98.6	85.0	78.3	97.0	82.5	77.4	91.6	84.8	78.6	95.6	85.3
UniFD (CVPR'23)	73.8	80.4	62.8	81.7	83.2	79.6	77.3	88.8	62.5	94.3	91.5	97.6	86.3	90.6	81.0	82.7
IID (CVPR'23)	83.8	86.6	79.9	92.8	89.5	97.0	79.1	72.2	94.6	80.7	72.1	100.0	80.5	71.9	100.0	83.4
FreqNet (AAAI'24)	52.2	51.2	94.4	54.5	52.3	100.0	71.2	66.7	84.5	77.5	69.8	96.6	78.8	70.8	98.0	66.8
ProDet (NeurIPS'24)	84.8	82.2	88.8	90.9	84.6	100.0	91.4	89.0	94.4	92.5	88.7	97.3	93.0	88.4	99.0	90.5
NPR (CVPR'24)	59.6	56.5	84.1	66.6	60.5	95.7	74.3	87.2	57.0	88.8	91.7	85.3	88.7	92.9	83.6	75.6
AIDE (ICLR'25)	58.9	58.7	59.6	78.6	70.3	99.0	84.5	92.4	75.2	86.8	93.1	99.0	93.0	92.7	93.3	80.4
Co-SPY (CVPR'25)	71.3	76.0	62.3	89.3	82.9	99.1	82.1	91.6	70.7	95.0	92.2	98.3	94.0	94.0	94.0	86.3
D^3 (CVPR'25)	72.5	70.3	77.8	82.1	74.7	96.9	91.1	91.7	90.3	96.3	94.2	98.6	94.7	91.3	98.6	87.3
Effort (ICML'25)	92.9	94.4	91.2	96.4	94.6	98.4	95.9	95.3	96.5	95.0	97.5	92.3	93.5	96.1	90.6	94.7
Qwen2.5-VL-7B	51.3	93.5	2.8	51.5	94.3	3.3	50.0	50.0	1.1	51.5	80.0	4.0	51.5	76.4	4.3	51.2
InternVL3-8B	55.9	55.5	59.0	53.7	53.6	54.6	56.3	83.8	15.6	52.0	66.6	8.0	52.3	69.4	8.3	54.0
MiMo-VL-7B	56.2	55.3	64.0	62.2	59.4	77.1	60.4	63.9	47.6	66.4	69.3	58.6	73.6	72.0	77.3	63.8
GLM-4.1V-9BThink	58.0	64.0	36.3	59.6	65.6	40.0	54.2	89.6	9.4	56.0	87.5	14.0	54.3	80.9	11.3	56.4
GPT-4o	49.2	20.0	0.5	53.0	84.2	8.0	49.8	33.3	0.5	63.2	100.0	26.5	52.5	91.6	5.5	53.5
Gemini-2.5-Pro	62.0	77.9	33.5	53.2	66.6	13.0	66.0	80.9	42.5	93.9	89.3	100.0	85.8	73.7	76.6	72.2
Veritas (ICLR'26)	90.9	90.2	91.7	96.1	92.7	100.0	99.8	99.6	100.0	99.8	99.6	100.0	100.0	100.0	100.0	<u>97.3</u>
COVE (base stage)	91.7	90.8	92.9	95.1	91.1	99.9	99.8	99.9	99.6	99.7	99.3	100.0	99.8	100.0	99.7	97.2
COVE (ours)	92.4	91.1	94.1	95.9	92.5	99.9	99.8	99.9	99.6	99.5	99.3	99.7	99.8	99.7	100.0	**97.5**

Table 8: Detailed cross-model results on HydraFake (%). Acc., P., and R. denote accuracy, precision, and recall, with fake as the positive class. Avg. is mean subset accuracy. The best averages are bold and the second best are underlined.

Method	ADF			FLUX			StarryAI			MAGI-1			HART			Infinity			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net (ECCV'20)	86.7	80.0	97.7	87.8	80.4	100.0	78.6	76.2	83.3	85.0	78.6	96.2	86.0	78.6	99.0	82.9	78.6	90.4	84.5
UniFD (CVPR'23)	90.7	92.9	88.0	93.8	91.7	96.3	82.5	88.8	74.3	73.0	84.7	56.1	94.4	92.2	97.0	90.7	92.8	88.2	87.5
IID (CVPR'23)	83.3	75.0	100.0	82.8	74.4	100.0	80.0	72.1	97.6	80.2	72.6	96.5	81.1	72.6	100.0	82.2	73.7	100.0	81.6
FreqNet (AAAI'24)	60.3	59.5	64.3	76.7	68.4	99.0	59.0	59.1	58.3	69.2	65.5	81.1	77.1	69.4	96.7	75.1	68.7	92.3	69.6
ProDet (NeurIPS'24)	92.6	88.3	98.3	94.2	89.5	100.0	88.2	88.0	88.3	91.9	87.5	97.7	93.8	89.0	99.8	93.1	88.7	98.6	92.3
NPR (CVPR'24)	68.8	86.0	45.0	91.2	91.5	90.6	59.5	78.8	26.0	82.6	91.9	71.5	91.3	93.1	89.3	84.0	92.6	73.8	79.6
AIDE (ICLR'25)	68.8	85.5	43.3	86.3	91.0	80.6	64.0	86.6	33.0	88.9	93.6	83.6	95.4	94.3	96.6	76.0	91.0	57.7	79.9
Co-SPY (CVPR'25)	93.5	93.6	93.3	95.5	94.1	97.0	85.3	93.4	76.0	93.3	92.2	72.7	96.6	94.5	98.9	95.3	94.1	96.7	93.3
D^3 (CVPR'25)	93.6	90.3	97.7	95.6	92.3	99.6	91.3	92.2	90.3	90.7	90.5	91.0	95.8	92.2	99.9	95.5	93.1	98.3	93.8
Effort (ICML'25)	82.8	97.1	67.6	96.5	96.0	97.0	78.0	94.2	59.6	90.5	96.5	84.1	97.8	97.2	98.5	98.3	97.3	99.4	90.7
Qwen2.5-VL-7B	50.0	50.0	1.3	50.0	50.0	1.3	49.7	0.0	0.0	50.0	50.0	0.9	52.0	83.8	4.9	52.9	87.3	6.8	50.8
InternVL3-8B	54.0	74.0	12.3	49.8	47.3	3.0	49.0	28.5	1.3	56.6	83.5	16.4	55.8	83.1	14.5	57.2	84.5	17.6	53.7
MiMo-VL-7B	74.5	74.7	74.0	77.1	76.7	78.0	82.5	68.0	55.3	60.3	64.0	47.1	82.4	77.9	90.6	81.4	77.2	89.2	76.4
GLM-4.1V-9BThink	55.2	82.9	13.0	52.3	76.9	6.6	50.5	66.6	2.0	51.6	81.5	4.2	68.4	96.1	38.2	60.7	95.5	22.5	56.5
GPT-4o	57.7	96.9	16.0	52.0	83.3	5.0	51.4	85.7	3.0	59.9	80.0	26.4	81.2	92.5	67.9	54.8	87.5	10.6	59.5
Gemini-2.5-Pro	64.9	78.8	41.2	92.4	90.3	94.9	82.8	92.3	72.0	62.5	79.0	34.1	93.4	88.5	100.0	93.2	89.1	98.5	81.5
Veritas (ICLR'26)	94.8	99.6	89.9	99.8	99.6	100.0	97.0	100.0	94.0	99.9	99.8	100.0	99.9	99.8	100.0	99.9	99.8	99.9	98.6
COVE (base stage)	96.0	99.3	92.7	99.7	99.3	100.0	99.3	99.7	99.0	99.8	99.6	100.0	99.7	99.7	99.8	99.8	99.7	100.0	<u>99.1</u>
COVE (ours)	98.5	97.7	99.3	99.8	99.7	100.0	99.5	100.0	99.0	99.6	99.4	99.8	99.9	99.8	100.0	99.8	99.7	99.8	**99.5**

Table 9: Detailed cross-forgery results on HydraFake (%). Acc., P., and R. denote accuracy, precision, and recall, with fake as the positive class. Avg. is mean subset accuracy. The best averages are bold and the second best are underlined.

Method	St.GANv2			ICLight			CodeF.			InfiniteY.			PuLID			FaceAda.			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net (ECCV'20)	41.3	20.7	6.2	48.9	47.9	24.8	71.9	72.8	69.8	84.9	78.2	96.8	85.5	79.5	95.7	72.6	77.5	62.8	67.5
UniFD (CVPR'23)	61.8	81.7	30.4	81.9	89.1	72.6	75.4	88.0	58.7	73.7	87.0	55.6	68.1	83.6	45.1	81.3	90.3	69.6	73.7
IID (CVPR'23)	41.4	32.8	16.6	53.3	54.3	41.0	79.7	72.9	94.4	81.8	73.4	99.6	81.8	73.3	99.9	73.7	68.2	87.1	68.6
FreqNet (AAAI'24)	33.1	11.4	5.0	73.1	67.6	88.9	70.3	66.6	81.5	72.8	67.4	88.3	77.4	69.9	96.2	67.7	64.9	75.0	65.7
ProDet (NeurIPS'24)	56.3	67.5	24.5	58.6	69.8	34.4	80.8	84.7	75.3	88.1	89.9	85.8	91.0	88.1	94.9	83.3	86.0	79.0	76.4
NPR (CVPR'24)	47.7	23.6	2.1	67.8	85.2	43.2	60.6	78.3	29.3	79.8	99.1	66.0	89.0	92.8	84.5	67.7	84.9	41.9	68.8
AIDE (ICLR'25)	56.7	74.6	20.3	79.2	91.5	64.3	86.1	90.8	80.2	74.2	89.4	55.0	62.4	83.9	30.5	75.7	90.3	56.7	72.4
Co-SPY (CVPR'25)	77.0	91.1	59.9	92.5	93.8	91.0	88.6	90.3	86.5	90.6	93.2	87.5	79.1	89.9	65.5	87.3	88.7	85.4	85.9
D^3 (CVPR'25)	62.4	82.4	31.5	71.6	86.3	51.4	82.9	90.9	73.1	80.0	88.7	68.7	82.4	89.8	73.1	73.7	84.1	57.4	75.5
Effort (ICML'25)	64.7	93.2	31.7	94.8	96.3	93.2	89.7	97.1	81.9	89.5	96.0	82.3	92.9	96.2	89.2	88.0	95.0	80.2	86.6
Qwen2.5-VL-7B	50.5	64.7	2.2	56.7	90.8	15.2	50.7	70.0	2.4	53.6	90.3	8.1	54.5	90.4	10.1	51.6	63.6	4.7	52.9
InternVL3-8B	62.9	87.7	30.0	54.2	77.1	12.0	62.9	86.9	30.5	63.6	92.1	29.7	54.8	80.1	12.7	67.7	96.3	35.8	61.0
MiMo-VL-7B	48.7	47.7	26.2	82.6	76.4	94.3	76.4	75.1	78.9	79.7	76.3	85.7	78.4	75.8	83.5	82.8	79.1	88.3	74.8
GLM-4.1V-9BThink	54.3	89.8	9.7	68.4	94.7	39.0	63.3	95.7	27.8	65.7	96.5	32.6	55.1	87.4	11.9	81.0	97.9	62.8	64.6
GPT-4o	66.4	82.6	41.5	58.9	100.0	18.0	52.5	91.6	5.5	64.4	100.0	29.0	60.9	94.0	23.5	55.5	100.0	10.1	59.8
Gemini-2.5-Pro	73.7	89.7	53.5	83.3	89.7	75.1	87.4	86.0	89.3	85.5	86.2	84.5	84.7	88.4	80.0	85.6	85.3	85.8	83.4
Veritas (ICLR'26)	90.3	99.5	81.0	75.7	99.5	51.5	97.0	98.6	95.3	91.8	98.9	84.5	95.1	99.5	90.6	91.7	98.6	84.6	90.3
COVE (base stage)	97.3	99.6	95.0	92.7	99.9	85.4	96.0	99.6	92.5	97.3	99.7	94.9	96.9	99.4	94.5	97.7	100.0	95.2	96.3
COVE (ours)	98.4	99.5	97.3	93.2	99.6	86.7	97.1	99.8	94.5	98.7	98.8	98.6	97.9	98.3	97.6	97.0	100.0	93.9	**97.1**

Table 10: Detailed cross-domain results on HydraFake (%). Acc., P., and R. denote accuracy, precision, and recall, with fake as the positive class. Avg. is mean subset accuracy. The best averages are bold and the second best are underlined.

Method	DeepFace.			InfiniteY.			Dreamina			HailuoAI			GPT-4o			FFIW			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net (ECCV'20)	57.7	54.7	89.1	78.5	72.6	91.6	55.6	55.0	60.9	68.6	63.3	88.6	66.2	63.6	75.5	66.4	64.8	71.4	65.5
UniFD (CVPR'23)	67.4	64.1	79.2	67.3	75.9	50.5	80.5	77.2	86.3	75.2	75.4	74.8	73.3	73.6	72.7	67.5	65.4	74.5	71.9
IID (CVPR'23)	65.2	65.1	65.5	69.9	63.0	95.9	63.8	58.0	99.6	63.3	57.6	100.0	63.8	58.2	97.8	64.2	66.6	56.6	65.0
FreqNet (AAAI'24)	50.6	50.3	98.7	67.0	62.3	85.8	62.1	58.4	83.4	59.3	56.9	76.8	58.3	56.3	73.3	51.2	50.6	95.0	58.1
ProDet (NeurIPS'24)	58.1	54.6	95.3	82.9	81.3	85.4	71.3	71.3	71.4	75.6	71.2	86.0	66.3	69.9	57.4	74.1	73.0	76.5	71.4
NPR (CVPR'24)	52.6	52.4	56.6	73.0	82.1	58.7	76.6	80.9	69.5	62.3	70.7	42.0	50.2	50.4	20.9	46.0	46.3	50.9	60.1
AIDE (ICLR'25)	59.7	57.5	73.9	67.9	75.2	53.2	49.7	49.4	25.2	58.0	60.6	45.8	51.9	53.0	33.3	59.2	57.3	71.2	57.7
Co-SPY (CVPR'25)	67.6	63.8	81.2	80.0	85.1	72.7	82.5	80.7	85.3	74.0	75.7	70.6	79.5	78.7	80.9	64.3	64.2	64.7	74.7
D^3 (CVPR'25)	69.7	70.0	69.1	74.6	78.8	67.2	78.1	72.2	91.4	70.9	67.7	79.8	80.8	75.1	92.0	64.3	66.0	58.7	73.1
Effort (ICML'25)	64.8	59.1	96.3	82.2	75.4	95.5	61.5	57.6	86.5	66.4	60.1	97.4	53.8	52.9	69.3	74.0	68.3	89.7	67.1
Qwen2.5-VL-7B	50.7	71.7	2.4	53.6	92.8	7.9	80.2	99.4	60.7	67.5	100.0	35.1	52.5	100.0	5.1	50.5	56.1	4.5	59.2
InternVL3-8B	54.4	55.5	44.1	67.1	83.6	42.7	77.1	85.4	65.3	66.5	78.9	45.0	47.4	37.5	7.6	51.8	52.0	45.1	60.7
MiMo-VL-7B	57.7	56.4	68.6	75.6	73.7	79.6	79.4	73.5	91.9	70.7	68.3	76.5	67.7	67.1	69.5	54.9	54.8	56.2	67.7
GLM-4.1V-9BThink	58.7	82.6	22.0	72.7	95.1	47.8	83.7	96.0	70.3	69.2	92.1	42.0	52.0	68.4	8.2	53.9	60.8	21.7	65.0
GPT-4o	49.4	41.7	2.5	62.0	98.0	24.5	90.7	98.8	82.4	73.7	100.0	47.5	58.0	94.4	17.0	52.8	59.3	17.8	64.4
Gemini-2.5-Pro	67.2	72.4	55.6	75.6	88.7	59.0	87.5	89.5	84.9	82.4	95.2	68.2	70.9	96.6	43.2	53.0	67.6	11.5	72.8
Veritas (ICLR'26)	58.6	54.7	100.0	84.1	94.1	72.8	92.3	90.0	95.1	90.2	86.5	95.2	89.2	86.3	93.2	78.5	76.1	83.0	82.2
COVE (base stage)	60.2	55.7	99.5	87.9	90.9	84.1	93.7	90.2	98.1	90.6	86.8	95.8	86.8	86.7	87.0	80.0	75.3	89.3	83.2
COVE (ours)	69.4	62.2	99.1	88.8	91.9	85.1	93.8	91.3	96.8	90.7	86.9	95.8	87.0	86.5	87.6	80.2	80.8	79.2	**85.0**

C DIRECT CLASSIFICATION OF EXPERT FEATURES

To test whether direct classification of expert features can reproduce the gains of COVE, we train classifiers on the frozen DINOv3, SigLIP2, and PiDiNet representations. All controls use the same HydraFake training and test splits as COVE. The pooled baselines use a linear head or an MLP with two layers. The spatial control uses four learned queries per expert to form twelve summary tokens, followed by a Transformer encoder with two layers and a binary head. These classifiers predict authenticity without generating a reasoning trajectory.

Table 11: Direct classification of frozen expert features on HydraFake (Acc., %). COVE (ours) is the final model.

Method	ID	CM	CF	CD	Overall
Expert pooled + Linear	60.3	57.4	55.1	52.3	55.2
Expert pooled + 2-layer MLP	60.7	57.5	56.7	52.5	55.8
Expert spatial Transformer	61.2	56.3	57.2	52.8	55.7
COVE (ours)	**97.5**	**99.5**	**97.1**	**85.0**	**94.0**

Table 11 shows that direct classification remains limited across all four groups. The MLP obtains the highest overall accuracy of 55.8%, followed by the spatial Transformer at 55.7% and the linear head at 55.2%. The spatial Transformer improves ID, CF, and CD over the MLP, but its lower CM accuracy offsets these gains. Thus, retaining spatial structure does not consistently improve generalization. COVE reaches 94.0% overall accuracy, exceeding the strongest direct classifier by 38.2 percentage points. The tested classifiers therefore fall well short of the complete system despite access to the same frozen experts.

The gap reflects the complete system, including its MLLM backbone and training objectives.

D EVALUATION OF REASONING QUALITY

Table 5 reports mean scores from GPT-4o and Gemini-2.5-Pro and Elo ratings for the seven compared models. Each scoring judge receives the image, its ground truth label, and the model response. The rubric assigns 1 to an incorrect verdict, 2 to correct but vague or repetitive reasoning, 3 to superficial analysis, 4 to correct and detailed reasoning grounded in the image, and 5 to comprehensive reasoning with fine visual evidence. Elo summarizes relative preference among these seven models. Comparisons with existing MLLM detectors are presented in Appendix G.

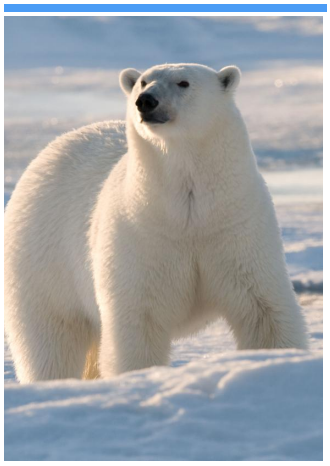
E PERFORMANCE ON BROADER BENCHMARKS

We use only facial forgery data for detection training and evaluate whether the learned detector generalizes to broader image generation and editing tasks. Table 12 reports accuracy and F1 on five benchmarks, with LOKI also evaluated on its facial subset. Despite the restricted training domain, COVE achieves 72.9% accuracy on LOKI, 85.8% on FakeClue, 71.2% on Forensics-Bench, and 89.2% on AIGIBench. It also reaches 86.1% on Nano-banana-150K, demonstrating transfer to recent image editing. Among the compared methods, COVE achieves the highest accuracy on LOKI, its facial subset, Forensics-Bench, and AIGIBench, while remaining competitive on FakeClue and Nano-banana. These results show that detection learned from facial forgeries can generalize to a wider range of visual content.

Table 12: Performance comparison on broader benchmarks (%). LOKI (facial) is the facial subset of LOKI. Nano-banana denotes Nano-banana-150K. Baseline results are from Tan et al. (2026). The best results are bold and the second best are underlined.

Method	**LOKI**		**LOKI (facial)**		**FakeClue**		**Forensics-Bench**		**AIGIBench**		**Nano-banana**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
UniFD	54.5	58.7	74.8	69.7	61.6	64.2	53.6	54.8	78.0	80.2	49.1	36.0
ProDet	53.8	56.6	63.2	66.4	62.9	69.8	65.1	72.0	74.8	73.9	64.6	63.8
Co-SPY	61.7	65.8	79.1	75.6	68.1	72.4	70.8	**76.0**	81.6	84.3	52.0	40.9
D^3	47.3	41.2	79.5	80.1	60.7	59.2	56.6	59.8	77.8	75.3	70.7	73.0
Effort	53.8	50.0	84.3	84.6	65.0	63.2	57.0	59.4	84.7	87.6	62.7	52.1
InternVL3-8B	53.0	51.6	52.6	15.3	59.1	62.2	60.5	65.4	55.6	56.5	51.3	38.9
MiMo-VL-7B	65.1	64.3	69.7	65.0	67.2	71.6	63.8	70.5	62.8	64.3	60.7	55.6
Veritas	72.1	77.8	89.0	88.2	**85.9**	**88.4**	70.8	74.9	88.9	90.4	**86.3**	**89.0**
COVE (ours)	**72.9**	**78.0**	**89.9**	**89.1**	85.8	88.0	**71.2**	75.4	**89.2**	**90.7**	86.1	88.8

Figures 4–7 present qualitative results on animals, documents, satellite imagery, and scenes from LOKI. Each figure contains a real image and a fake image together with the model's responses. The correct verdicts in these examples illustrate how the detector applies beyond the facial domain used for training.



Real Image from LOKI, animal

<fast> At first glance, the image appears authentic with no obvious signs of manipulation. The polar bear and its surroundings look natural and coherent. </fast>
<planning> To determine the authenticity more confidently, we should examine the image through three progressive layers: first checking for visual structural abnormalities, then inspecting for subtle generator artifacts, and finally evaluating for violations of physical laws. This tiered approach ensures a comprehensive and logical verification process.</planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships: <vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency: <vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns: <vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> The image shows no structural inconsistencies; the polar bear's proportions are realistic, its posture natural, and its expression neutral. There are no mismatches in color or symmetry, and the features align with typical feline anatomy. Upon closer inspection, the fur texture appears realistic with individual strands visible, and there are no signs of artificial edge smoothing, blending boundaries, or unnatural blurring. The lighting and shadows are consistent with a natural light source from above, and the background matches the subject in style and illumination. There is no evidence of stylization or focus discrepancy that would suggest digital generation. </reasoning>
<conclusion> Based on the absence of visual anomalies, subtle artifacts, or physical inconsistencies, the image is most likely authentic. While not perfectly clear in judgment, the cumulative evidence strongly supports its genuineness.</conclusion>
<answer> real </answer>



Fake Image from LOKI, animal

<fast> At first glance, the image appears realistic, with no glaring signs of artificial generation. </fast>
<planning> Given the initial impression of realism, a thorough and systematic examination is essential to discern authenticity. This involves evaluating structural integrity, identifying subtle generation artifacts, and assessing any inconsistencies with physical realism, particularly in texture and focus. This comprehensive approach ensures that even the more nuanced indicators of AI generation are considered. </planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships: <vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency: <vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns: <vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> Initially, the puppy's structure and proportions seem consistent with a real photograph. The grassy background is appropriately integrated, and there are no obvious distortions in the anatomy of the puppy. The overall focus of the image is sharp, which is typically acceptable in real photography. However, upon closer inspection, the skin texture of the puppy, particularly on the sides and back, appears unusually smooth. This smoothness lacks the natural variation and fine detail typically seen in real dog fur, which can be a sign of digital generation. Additionally, there is a slight blurriness around the edges of the legs and tail, which detracts from the image's overall sharpness and realism. These subtle inconsistencies suggest manipulation. The lighting on the puppy seems mostly consistent, but it lacks the nuanced texture that would be expected in a naturally lit photograph. </reasoning>
<reflection> Wait a moment—there's something else that stands out upon further reflection. The smoothness of the skin, while subtle, is accompanied by a lack of high-frequency detail, which is a known limitation of some AI generation models. Real fur usually has tiny, visible strands and natural imperfections, which are missing here. This absence, combined with the slight blurriness, suggests that the image might be designed to mimic real photography but falls short upon closer inspection. These elements together point towards a sophisticated deepfake that attempts to pass as real but shows minor signs of its artificial origin. </reflection>
<conclusion> Despite its convincing appearance at first glance, the image's overly smooth skin texture, slight blurriness on dynamic parts, and lack of fine detail suggest it is AI-generated. These subtle yet telling signs, combined with the overall realism but slight inconsistencies in physical plausibility, lead to the conclusion that the image is medium hard to detect as a deepfake. Careful scrutiny reveals the artificial nature beneath the realistic facade. </conclusion>
<answer> fake </answer>

Figure 4: Qualitative results on animal images from LOKI. The top and bottom examples show a real image and a fake image, respectively, together with the responses of COVE.



Real Image from LOKI, document

<fast> At first glance, the image appears authentic with no obvious signs of manipulation. The content looks coherent and naturally presented. </fast>
<planning> To determine the authenticity, we should examine the image through a progressive lens: starting with visually apparent structural features, then moving to more subtle generator-related artifacts, and finally assessing for violations of physical plausibility. Given the initial impression and the absence of clear anomalies, a structured forensic approach is appropriate. </planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships: <vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency: <vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns: <vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> The image shows no structural abnormalities such as mismatched color tones, distorted proportions, or unnatural expressions. Facial features are symmetric and follow biometric norms, and the overall composition is consistent with real human anatomy. Upon closer inspection, there are no jagged edges, unnatural textures, or blending artifacts that might suggest digital generation. The image maintains natural skin texture, appropriate blurring, and fine detail, all of which support its realism. Even under deeper scrutiny, the lighting and shadows align with real-world physics, and there are no contextual mismatches in makeup, background, or focus that would imply manipulation or artificial enhancement. </reasoning>
<conclusion> Synthesizing all the evidence, the image does not exhibit any signs of being fake. It maintains structural integrity, natural appearance, and physical plausibility, leading to the conclusion that it is authentic. </conclusion>
<answer> real </answer>



Fake Image from LOKI, document

<fast> At first glance, the image appears realistic, with no immediate signs of forgery that stand out. </fast>
<planning> Given the initial impression of realism, a thorough and systematic examination is necessary. This involves checking for structural coherence, identifying subtle visual artifacts, and ensuring that the image adheres to physical and logical realism, especially in terms of text and layout. This comprehensive approach is essential to uncover any hidden signs of manipulation. </planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships: <vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency: <vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns: <vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> Upon closer inspection, several issues become apparent. The text alignment is inconsistent, particularly in the first paragraph where the text is misaligned and does not follow the expected flow. This suggests a lack of careful composition, which is unusual in professionally designed documents. Additionally, the captions for the images are problematic: the caption for the building image is missing entirely, and the caption for the person is awkwardly placed and incomplete. This is a significant oversight, as captions are typically essential for context and understanding in such layouts. Furthermore, the Chinese text on the right side of the page is not only misaligned but also seems to be cut off, with only partial sentences visible. This kind of disruption is not typical in well-prepared documents and indicates potential editing errors. The mathematical equations and explanations are also somewhat jumbled, with the text not flowing smoothly from one point to the next, which detracts from the clarity and professionalism of the document. </reasoning>
<reflection> Wait a moment—there's something else that stands out upon further reflection. The overall layout seems slightly off, with elements not quite fitting together seamlessly. The images and text appear to be placed in a way that prioritizes fit over function, leading to a disjointed visual experience. This kind of mismatch is often a telltale sign of AI generation, where the system struggles to maintain a cohesive and natural layout when synthesizing complex elements like text and images together.</reflection>
<conclusion> While the image initially projects a sense of authenticity, the alignment issues, missing and misplaced captions, and disrupted text flow strongly suggest that it is AI-generated. These elements, though subtle, collectively undermine the realism of the image, indicating that it is a medium-hard forgery to detect without careful scrutiny.</conclusion>
<answer> fake </answer>

Figure 5: Qualitative results on document images from LOKI. The top and bottom examples show a real image and a fake image, respectively, together with the responses of COVE.



Real Image from LOKI, satellite

<fast> At first glance, the image appears authentic with no obvious signs of manipulation. The aerial view of a building complex looks consistent with real-world photography. </fast>
<planning> To verify authenticity, we proceed in a structured manner: first examining for obvious structural anomalies, then inspecting for subtle generator artifacts, and finally analyzing for violations of physical plausibility. This progressive approach ensures that both surface-level and deeper inconsistencies are considered.</planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships: <vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency: <vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns: <vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> The image shows no visual structural abnormalities such as color mismatches, distorted proportions, or unnatural expressions. The buildings, roofs, and surrounding elements are proportionate and aligned realistically. Moving beyond the surface, subtle artifacts often introduced by generative models—like edge irregularities, texture anomalies, or unnatural blurring—are absent. The image maintains natural edges, realistic texture, and clear detail across all regions. Even under deeper scrutiny, it adheres to physical laws: lighting and shadows are consistent, there are no implausible focus shifts, and the overall style remains documentary without artificial stylization. Although the image quality is not optimal, no misleading artifacts are present that would suggest manipulation.
</reasoning>
<conclusion> Synthesizing all evidence, the image shows no signs of being fake. It maintains structural integrity, lacks generator-specific artifacts, and conforms to physical realism. Therefore, it can be judged as authentic with high confidence. </conclusion>
<answer> real </answer>



Fake Image from LOKI, satellite

<fast> At first glance, the image appears to be a realistic aerial view of a landscape, with no immediately obvious signs of manipulation. </fast>
<planning> Given the lack of clear visual anomalies, a structured forensic analysis is necessary. We begin by examining the image for structural coherence, then move into subtle generator artifacts, and finally assess for violations of physical plausibility. This multi-layered approach is essential for detecting the nuanced signs of AI generation in a highly realistic image. </planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships: <vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency: <vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns: <vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> The image presents a visually consistent landscape with no apparent structural abnormalities. The colors are uniform, the lighting is even, and the overall composition adheres to realistic geographical and topographical norms. There are no distortions in features or mismatches in lighting direction that would immediately suggest manipulation. To dive deeper, we examine potential artifacts introduced by generative models. The edges are smooth, the textures are natural, and there are no visible blending boundaries or unnatural blurs. However, upon close inspection, the image lacks high-frequency details--subtle variations in texture and fine spatial relationships between elements that are typically present in real-world aerial photography. This absence is a known artifact in AI-generated images, which often struggle to replicate the full complexity of natural scenes at a microscopic level. Considering physical plausibility, the lighting and shadows align logically, and there are no contextual mismatches or implausible elements. The image maintains a realistic style without overt stylization or focus inconsistencies. Yet, the slight lack of detail suggests a level of artificial smoothing, which, while not immediately apparent, becomes evident under detailed scrutiny. </reasoning>
<reflection> Wait--what if the very realism of the image is part of the deception? The absence of flaws might itself be a flaw. Real aerial images often contain minor imperfections or variations that AI might over-smooth to maintain an "idealized" version of reality. This subtle over-smoothing, while not a direct artifact, could contribute to the image's uncanny realism and make it harder to detect as fake. </reflection>
<conclusion> While the image passes initial and even intermediate scrutiny with flying colors, the lack of high-frequency detail and the potential over-smoothing of textures suggest AI generation. These subtleties, combined with the image's unnatural lack of flaws, make it a medium-hard forgery to detect. The realism is convincing, but the underlying generative reveal its artificial origin upon detailed inspection.</conclusion>
<answer> fake </answer>

Figure 6: Qualitative results on satellite images from LOKI. The top and bottom examples show a real image and a fake image, respectively, together with the responses of COVE.



Real Image from LOKI, scene

<fast> At first glance, the image appears authentic with no obvious signs of manipulation. </fast>
<planning> To confirm authenticity, the analysis will proceed in three stages: first checking for visible structural inconsistencies, then inspecting for subtle generator artifacts, and finally evaluating adherence to physical laws. Since the initial judgment is already positive, this structured approach allows for thorough yet logical verification. </planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships:
<vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency:
<vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns:
<vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> The image shows no visual structural abnormalities such as color mismatches, distorted proportions, or unnatural expressions. The buildings, roofs, and surrounding elements are proportionate and aligned realistically. Moving beyond the surface, subtle artifacts often introduced by generative models—like edge irregularities, texture anomalies, or unnatural blurring—are absent. The image maintains natural edges, realistic texture, and clear detail across all regions. Even under deeper scrutiny, it adheres to physical laws: lighting and shadows are consistent, there are no implausible focus shifts, and the overall style remains documentary without artificial stylization. Although the image quality is not optimal, no misleading artifacts are present that would suggest manipulation. </reasoning>
<conclusion> All levels of analysis confirm the image is authentic. It exhibits no signs of digital manipulation or synthetic generation and aligns with the characteristics of a real photograph. </conclusion>
<answer> real </answer>



Fake Image from LOKI, scene

<fast> At first glance, the image appears convincingly realistic, with a serene and well-composed scene of a sunset over a waterfront community. </fast>
<planning> Given the initial impression of realism, a thorough and structured examination is essential. This involves assessing structural integrity, identifying subtle generation artifacts, and verifying physical plausibility, particularly in lighting and texture. This methodical approach is necessary to uncover the less obvious signs of AI manipulation. </planning>
<reasoning> Structural feature analysis reveals the global and local spatial relationships:
<vis_dino_0><vis_dino_1><vis_dino_2><vis_dino_3> Semantic alignment analysis indicates text-image consistency:
<vis_siglip_0><vis_siglip_1><vis_siglip_2><vis_siglip_3> Edge and boundary detection shows artifact patterns:
<vis_pidinet_0><vis_pidinet_1><vis_pidinet_2><vis_pidinet_3> Upon closer inspection, the structural elements of the houses and landscape maintain coherence and proportion, showing no immediate signs of distortion or inconsistency. However, when focusing on the details such as the rooftops, I notice there is a peculiar bright spot on one of the rooftops. This spot doesn't interact naturally with the light source from the sun, suggesting a synthetic overlay or reflection that doesn't conform to real-world lighting conditions. Furthermore, the texture of the palm leaves appears slightly unnatural, with certain fronds exhibiting unusual highlights and shadows that don't match typical foliage behavior. \n\nAI-generated images often struggle with accurately representing texture and reflective properties, especially when blending different materials and natural elements like plants. Additionally, the overall scene possesses a somewhat stylized quality, with the brightness and saturation of the sunset appearing exaggerated. While subtle, these elements contribute to an artificial ambiance that diverges from natural scenery. </reasoning>
<conclusion> Although the image presents a convincing initial view of realism, the abnormal bright spot on the rooftop and the unnatural texture of palm leaves indicate AI manipulation. These subtle inconsistencies, combined with the image's stylized lighting, suggest that it is likely AI-generated. This line of reasoning, involving detailed observation and logical deduction, helps reveal the image as a medium-hard forgery, requiring nuanced detection to identify its artificial nature. </conclusion>
<answer> fake </answer>

Figure 7: Qualitative results on scene images from LOKI. The top and bottom examples show a real image and a fake image, respectively, together with the responses of COVE.

F ANALYSIS OF HYPERPARAMETERS

We examine the KL penalty coefficient β' and group size G in GRPO. All variants start from the same DPO checkpoint and are evaluated before contrastive forensic grounding. We vary one parameter at a time while keeping the training images, update count, reward weights, and remaining settings fixed. Gray rows mark the default configuration.

Table 13: Effect of β' in GRPO (Acc., %).

Value of β'	ID	CM	CF	CD
$\beta' = 0.04$	96.8	98.6	95.7	81.5
$\beta' = 0.01$	96.9	98.6	96.1	82.4
$\beta' = 0.001$	97.1	98.8	96.0	82.7
$\beta' = 0$	97.2	99.1	96.3	83.2

Table 14: Effect of G in GRPO (Acc., %).

Value of G	ID	CM	CF	CD
$G = 4$	97.2	99.1	96.3	83.2
$G = 8$	97.3	98.9	95.1	83.0
$G = 12$	97.0	98.1	94.1	81.4
$G = 16$	96.9	97.6	94.1	81.7

KL penalty coefficient β'. The coefficient β' controls how strongly the policy is penalized for deviating from the DPO reference. Table 13 examines this constraint with $G = 4$. Removing the penalty gives the best results in all four groups. Relative to $\beta' = 0.04$, CF and CD accuracy increase by 0.6 and 1.7 percentage points, respectively, while ID and CM also improve. This suggests that, once SFT and DPO establish structured reasoning, a strong reference constraint can restrict adaptation to the combined rewards. Allowing the policy to move beyond its initial behavior is particularly useful under domain shift. We therefore use $\beta' = 0$.

Group size G. The group size determines how many trajectories contribute to the relative advantage estimate for each image. With $\beta' = 0$, $G = 4$ gives the best CM, CF, and CD accuracy (Table 14). Increasing G to 8 gains only 0.1 percentage points on ID and lowers CF and CD by 1.2 and 0.2 points. Groups of 12 or 16 further underperform $G = 4$ across all four settings. A plausible explanation is that the trajectories share the same visual evidence and a policy already trained by SFT and DPO, so extra samples can repeat similar analyses without providing better decision cues. Larger groups also require more generated trajectories at each update. We therefore use $G = 4$ for stronger generalization with fewer rollouts.

G QUALITATIVE COMPARISONS WITH EXISTING MLLM DETECTORS

We compare the reasoning outputs of COVE with M2F2-Det (Guo et al., 2025), FFAA (Huang et al., 2024), FakeShield (Xu et al., 2025), SIDA-13B-description (Huang et al., 2025), FakeVLM (Wen et al., 2025), and Veritas (Tan et al., 2026) on the same images. Figures 8–10 cover face relighting, entire face generation, and face swapping. The comparison examines both verdict correctness and how each response connects visual observations to its decision.

M2F2-Det and FFAA classify all three displayed forgeries as real, relying on features such as natural skin, wrinkles, or coherent facial structure. These observations describe a plausible appearance but do not establish authenticity. FakeShield detects the InfiniteYou and FFIW examples but misses the relit face. SIDA-13B-description flags the first two examples as partially tampered and accepts the FFIW example as real. Its description of local tampering in the InfiniteYou example also differs from the generation of the entire face. FakeVLM gives the correct verdict in all three cases but repeats the same generic explanation without identifying the alleged inconsistencies.

Veritas correctly detects the FFIW swap but judges the relighting and generation examples as real. Its structured responses in these two cases accept the natural appearance without finding evidence of manipulation. COVE revisits this initial impression and predicts fake in all three examples. For IC-Light, its response discusses lighting differences across the face, hair, and background. For InfiniteYou, it examines focus around the ear and hair, skin texture near the nose and mouth, and facial shadows. For FFIW, it points to changes in skin tone and texture around the nose and mouth, then revisits the transition between the hair and background during reflection. These responses make the image regions and visual properties motivating the final decision explicit.



Figure 8: Qualitative comparison on an IC-Light image with face relighting. COVE relates its fake verdict to lighting differences across the face, hair, and background.



Figure 9: Qualitative comparison on a face image generated by InfiniteYou. COVE examines local texture, focus, and shadow cues before predicting fake.



Figure 10: Qualitative comparison on an FFIW image with face swapping. Both COVE and Veritas predict fake and describe local inconsistencies in texture, lighting, and facial boundaries.