SEE, THEN REASON: REFLECTIVE ON-POLICY SELF-DISTILLATION FOR DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have enabled deepfake detectors to support forensic reasoning alongside binary real-or-fake predictions. However, it remains unclear whether current MLLMs can preserve and effectively exploit the visual information required for deepfake detection. Our analysis reveals that deepfake detection performance consistently degrades across successive language-model layers. Reasoning-based SFT also underperforms the strongest intermediate-layer probe and exhibits a larger real-fake accuracy gap, indicating that the detection capability available in intermediate visual representations is not fully preserved in the final prediction. To address these limitations, we propose **SeeTR (See, Then Reason)**. Instead of entangling visual judgment with reasoning, SeeTR grounds the initial judgment in the most discriminative intermediate representation. It then builds reflective reasoning on this strong visual detection capability to determine whether the judgment should be kept or revised. Building on this reflective structure, we introduce **Contrastive Teacher On-Policy Self-Distillation (CT-OPSD)** to refine the model's reflective behavior on its own rollouts. CT-OPSD forms contrastive teacher contexts conditioned on the ground-truth label and its opposite, using their token-level disagreement to strengthen the model's ability to keep correct initial judgments and revise incorrect ones. Under the cross-domain & forgery setting, the initial visual judgment alone already outperforms prior MLLM-based detectors, and SeeTR further improves this visual baseline through reflective reasoning, achieving the highest average balanced accuracy among the evaluated MLLM-based detectors.

1 INTRODUCTION

Recent generative models can create and manipulate facial images with a level of realism that makes visual verification increasingly difficult. Their misuse for impersonation (Ciftci et al., 2020; Bohacek & Farid, 2022), misinformation (Vaccari & Chadwick, 2020; Dobber et al., 2021), and identity fraud (Kaushik et al., 2025; Bateman, 2022) has made facial deepfake detection an important security problem. Deepfake detection therefore depends on recognizing diverse manipulation traces and generalizing to unseen generation methods. Existing vision-based detectors analyze frequency artifacts (Woo et al., 2022; Gu et al., 2022a; Tan et al., 2024), spatio-temporal inconsistencies (Zheng et al., 2021; Gu et al., 2022b; Yan et al., 2025b), and biological signals (Wu et al., 2024; Birla et al., 2025) for real-or-fake classification. However, these approaches typically provide only a binary prediction, offering limited insight into the forensic evidence supporting their decisions. This limitation has motivated recent work on MLLM-based deepfake detection, which couples authenticity prediction with forensic reasoning.

Existing MLLM-based studies typically construct forensic rationale data using human annotators or external models and apply supervised fine-tuning (SFT) to train forensic reasoning (Zhang et al., 2024; Chen et al., 2026; Wen et al., 2026; Tan et al., 2026a). More recent approaches further incorporate preference optimization or reinforcement learning to improve reasoning quality and generalization (Tan et al., 2026b;a). However, how effectively these models preserve and exploit visual information for deepfake detection remains underexplored. Our analysis in Figure 1 reveals two key limitations. ***(i)* Layer-wise detection degradation:** layer-wise probing shows that detection performance generally declines toward later language-model layers; ***(ii)* Detection degradation and class skew under SFT:** reasoning-based SFT underperforms the strongest intermediate-layer probe

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

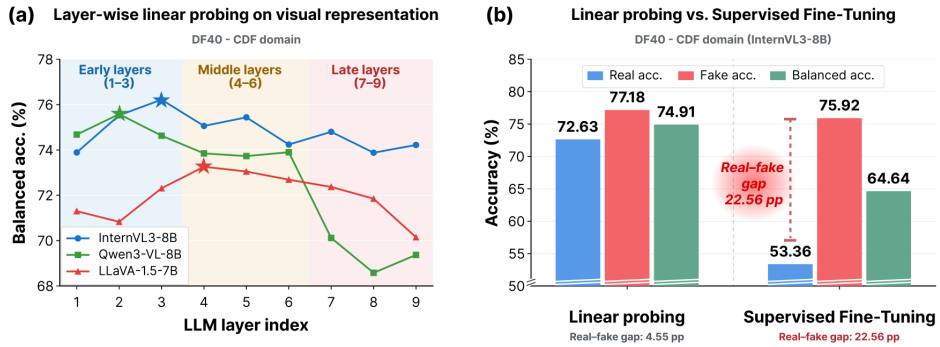

Figure 1: **Visual detection performance across LM layers and after reasoning-based SFT.**[1] (a) **Layer-wise probing**: Linear-probing performance generally decreases toward later LM layers for InternVL3, Qwen3-VL, and LLaVA-1.5-7B. (b) **Detection degradation and class skew under SFT**: Compared with the strongest intermediate-layer probe, reasoning-based SFT achieves lower balanced accuracy and exhibits a larger real–fake accuracy gap, and is skewed toward the fake class.

and exhibits a larger real-fake accuracy gap, indicating that its final prediction does not preserve the probe's detection performance or class balance. These findings motivate establishing a strong initial visual judgment from the most discriminative intermediate representation before reasoning. However, the initial visual judgment can still be incorrect, motivating subsequent reflection to determine when it should be revised. The central challenge is therefore to learn effective *keep-or-revise* behavior: keeping correct initial judgments while revising incorrect ones. Our experiments show that this behavior consistently improves upon the initial visual judgment under distribution shifts.

To address these limitations, we propose **SeeTR** (See, Then Reason). Rather than entangling visual judgment with reasoning, SeeTR forms an initial visual judgment using a detector trained on the most discriminative intermediate MLLM representation and uses subsequent reflection to determine whether it should be kept or revised. Training proceeds in two stages. In the first stage, the model learns this keep-or-revise behavior through visual supervision and policy-aligned reflective responses, allowing reflection to build on the initial judgment rather than override it. Building on this behavior, we introduce **Contrastive Teacher On-Policy Self-Distillation (CT-OPSD)** to further refine reflection on the model's own rollouts. CT-OPSD extends *on-policy self-distillation* (Zhao et al., 2026; Shenfeld et al., 2026) with positive and negative teacher contexts conditioned on the ground-truth label and its opposite, respectively. The same teacher processes the student's on-policy reflective rollout under both contexts. By emphasizing the positive teacher's supervision at reasoning positions with larger token-level disagreement, CT-OPSD helps the model strengthen its reflective reasoning to confirm correct initial visual judgments and revise incorrect ones. Our analysis shows that CT-OPSD achieves superior detection performance compared with standard OPSD.

In cross-domain & forgery evaluation, the initial visual judgment achieves 74.27% BAcc, and reflective reasoning further raises it to 77.18% (+2.91 points), establishing state-of-the-art performance among MLLM-based detectors. Recent MLLM-based detectors also exhibit substantial class skew under domain shift, consistent with our analysis in Figure 1. Across five unseen real-source domains, reflection improves mean BAcc from 72.04% to 75.78%.

2 Related Work

Facial deepfake detection. Recent studies on facial deepfake detection have shown that foundation models such as CLIP (Radford et al., 2021) provide strong transferable representations for generalizable detection. UnivFD (Ojha et al., 2023) first demonstrated that frozen CLIP representations with simple probing can generalize well across unseen generators. Subsequent studies improved this capability by exploiting intermediate visual features (Yermakov et al., 2025; Jung et al., 2026) and preserving pretrained knowledge during adaptation. Khan & Dang-Nguyen (2024) showed that

[1]Both analyses use the FF++ training domain of DF40 (Yan et al., 2024) and are evaluated on the unseen CDF domain under the cross-domain setting. Experimental details are provided in Appendix A.3.

lightweight adaptation can outperform full fine-tuning, while Effort (Yan et al., 2025a) preserves the principal weight subspace and adapts the residual space for forgery-specific knowledge. More recently, SD^2 (Jung et al., 2026) extends CLIP-based detection to richer multimodal knowledge by incorporating fine-grained forgery labels and image descriptions. However, as generative models continue to grow in number and diversity, these approaches remain largely classification-oriented and provide limited evidence for explaining why an image is predicted as real or fake.

To move beyond binary predictions, recent studies have explored MLLMs for explainable deepfake detection. Zhang et al. (2024) introduce human-annotated VQA supervision to jointly predict authenticity and verbalize facial forensic cues. X^2-DFD (Chen et al., 2026) constructs capability-aware forensic rationales using external MLLMs and specialized detectors to better match the visual capability of the target MLLM. SIDA (Huang et al., 2025) extends the task to joint detection, localization, and explanation. More recent studies adopt an SFT-to-RL training recipe to improve reasoning and generalization. Veritas (Tan et al., 2026b) constructs externally generated, pattern-aware CoTs that mimic a human-like reflective process, followed by preference optimization and reinforcement learning. Veritas++ (Tan et al., 2026a) follows a similar cold-start paradigm, using external MLLMs to restructure human-verified perceptual evidence into structured reasoning trajectories before perception-oriented RL and on-policy distillation. However, both approaches rely on externally constructed reasoning trajectories for cold-start supervision. Prior studies show that such off-policy supervision can mismatch the target model's generation distribution and impair generalization (Agarwal et al., 2024; Bao et al., 2025; Mendes et al., 2026). Separately, our analysis in Figure 1 shows that strong intermediate-layer detection performance is not fully preserved after reasoning-based SFT. In contrast, SeeTR establishes the initial judgment from the most discriminative intermediate representation and trains subsequent reflection using target-model-generated responses and on-policy rollouts.

On-policy self-distillation. Recent methods based on reinforcement learning with verifiable rewards (RLVR) have substantially improved the reasoning capability of large language models by optimizing model-generated trajectories with verifiable feedback (Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025). These approaches benefit from on-policy exploration, allowing the model to improve directly from responses sampled from its current policy. However, verifiable rewards are typically assigned at the response level and therefore provide limited guidance about which intermediate reasoning steps should be reinforced or corrected (Yang et al., 2026). This motivates learning signals that preserve on-policy training while providing finer-grained supervision throughout the generated trajectory.

On-Policy Self-Distillation (OPSD) (Zhao et al., 2026; Shenfeld et al., 2026) provides token-level supervision by distilling student-generated rollouts from a privileged teacher conditioned on additional information such as ground-truth labels. Recent studies extend OPSD along two main directions. One line enriches the privileged teacher with additional visual or task-specific information (Yuan et al., 2026; Tian et al., 2026; Aniri et al., 2026), while another assigns non-uniform importance to token positions based on uncertainty, teacher–student disagreement, or positional reliability (Xu et al., 2026; Ke et al., 2026; Liu et al., 2026). Building on these developments, our approach introduces contrasting privileged contexts using the ground-truth label and its opposite, and uses their disagreement to reweight token-level distillation for deepfake reasoning.

3 METHOD

3.1 OVERVIEW OF SEETR

Existing MLLM-based deepfake detectors typically generate forensic reasoning and the final authenticity prediction within a single reasoning process. In contrast, SeeTR introduces an initial visual judgment before reasoning, forming a three-stage reflective response:

$$y = \underbrace{\langle \texttt{first} \rangle \hat{c}_0 \langle / \texttt{first} \rangle}_{\text{Initial Visual Judgment}} \; \underbrace{\langle \texttt{think} \rangle r \langle / \texttt{think} \rangle}_{\text{Reflective Reasoning}} \; \underbrace{\langle \texttt{answer} \rangle \hat{c}_1 \langle / \texttt{answer} \rangle}_{\text{Final Decision}} \tag{1}$$

where $\hat{c}0$ denotes the initial visual judgment, r the reflective forensic reasoning, and $\hat{c}1$ the final decision. This structure changes the role of reasoning from directly producing a real-or-fake decision to verifying an initial visual judgment and deciding whether to keep or revise it.

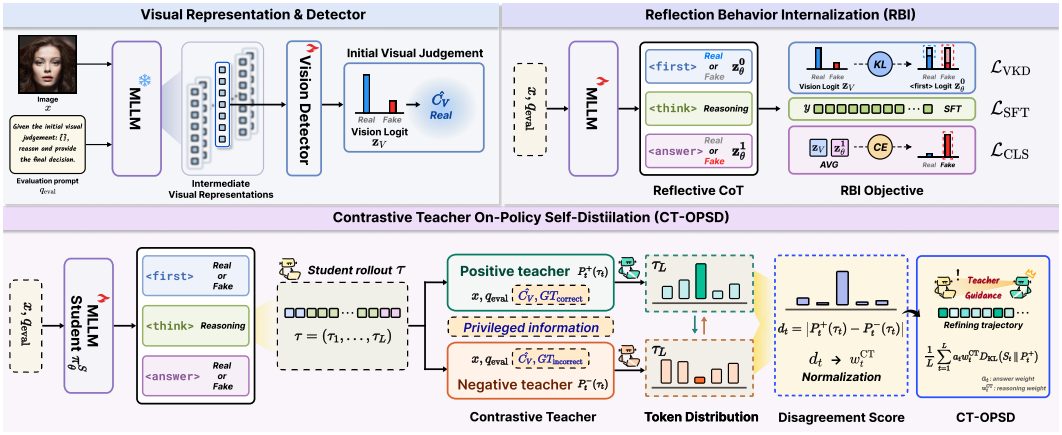

Figure 2: **Overview of SeeTR. Top-left:** We select the intermediate visual representation with the strongest deepfake detection performance and train a lightweight visual detector whose prediction serves as the initial visual judgment. **Top-right:** The model internalizes the reflective response from the initial judgment to the final decision, guided by the visual-detector logits together with autoregressive and classification supervision. **Bottom:** CT-OPSD further refines forensic reflective reasoning on the student's own on-policy rollout by using positive–negative teacher disagreement to emphasize informative reasoning positions during distillation.

Training proceeds in two stages. The first stage (Section 3.2) trains each reflection model to internalize this keep-or-flip response structure, grounding the initial visual judgment in the visual detector and supervising the subsequent reasoning and final decision. We then introduce **Contrastive Teacher On-Policy Self-Distillation (CT-OPSD)** (Section 3.3) to further refine reflective reasoning on each model's own trajectories. CT-OPSD conditions the same teacher on the ground-truth label and its opposite, and uses their token-level disagreement to emphasize reasoning positions where the two contexts diverge during distillation. Both stages update only LM LoRA adapters; the vision encoder and detector stay frozen.

3.2 Reflective Behavior Internalization

Before CT-OPSD, we first train the model to internalize a reflective behavior (*See, then reason*). We refer to this stage as **Reflective Behavior Internalization (RBI)**. Rather than simply applying autoregressive SFT to the reflective responses, RBI incorporates visual supervision from the visual-detector logits. It distills the detector distribution into the initial visual judgment and reuses the detector logits to regularize the final decision, encouraging a more decisive choice between confirming and flipping the initial judgment. Autoregressive supervision over the full response simultaneously teaches the reflective reasoning trajectory used for subsequent on-policy refinement.

Visual judgment from intermediate representations. Motivated by the visual capability degradation observed in Figure 1(a), we train a visual detector g on the intermediate MLLM representation with the highest probing performance. Given the pooled visual representation $h_V(x, q_{\text{eval}})$,[1] the detector produces the visual logits $\mathbf{z}_V = g(h_V(x, q_{\text{eval}})) = [z_V^{\text{real}}, z_V^{\text{fake}}]$. The higher-scoring class defines the initial visual judgment $\hat{c}_0 \in \{\text{real}, \text{fake}\}$, and $\mathbf{z}_V$ is used as visual supervision during training. Details of detector construction and the extraction of h_V are provided in Appendix A.3.

Policy-aligned dataset construction. Existing MLLM-based facial deepfake detectors commonly learn from reasoning annotations authored by human annotators or external models (Zhang et al., 2024; Chen et al., 2026; Huang et al., 2025; Tan et al., 2026b;a). Such externally constructed annotations may not align with the target model's generation distribution, potentially limiting generalization (Agarwal et al., 2024; Bao et al., 2025; Mendes et al., 2026).

[1] q_{eval} is the shared prompt used for visual-detector training and evaluation, as well as MLLM training and inference, with its full template provided in Figure 7.

To reduce this mismatch, we restrict the external MLLM to forensic artifact extraction, while the target MLLM generates the reflective response. To guide the extraction, we curate a forensic taxonomy of 59 artifact patterns by reviewing prior deepfake-forensics literature and manually inspecting samples across major forgery types. For manipulated images, the external MLLM identifies image-specific forensic evidence from the corresponding taxonomy. The target MLLM is conditioned on the visual-detector prediction $\hat{c}_V$ and receives the ground-truth label c, together with the extracted evidence when available, as privileged information. The privileged prompt presents $\hat{c}_V$ as the initial visual judgment and specifies whether it should be kept ($\hat{c}_V = c$) or flipped ($\hat{c}_V \neq c$). The target MLLM then generates the reflective response y using the structured format defined in Equation 1.

For first-stage training, the privileged context is removed, and the evaluation prompt q_{eval} is used as the model input. We construct the policy-aligned training dataset as

$$\mathcal{D}_{\text{RBI}} = \{(x_i, q_{\text{eval}}, y_i)\}_{i=1}^{N}, \tag{2}$$

where y_i is generated by the target MLLM under the privileged context. This construction keeps the reasoning supervision closer to the target model's generation behavior while guiding reflection with image-specific forensic evidence and the ground-truth label. Details of the forensic taxonomy, prompts, and filtering procedures are provided in Appendix B.

Visual knowledge distillation for initial visual judgment. To ground the initial visual judgment, we distill the visual-detector distribution into the MLLM before reflective reasoning. At the `Initial Visual Judgment` token position, the `real` and `fake` logits selected from the MLLM vocabulary form $\mathbf{z}_\theta^0 = [z_{\theta,\text{real}}^0, z_{\theta,\text{fake}}^0]$. The corresponding class distributions $\mathbf{p}_\theta^0$ and $\mathbf{p}_V$ are obtained by applying softmax to $\mathbf{z}_\theta^0/T_{\text{LM}}$ and $\mathbf{z}_V(x, q_{\text{eval}})$, respectively. The visual knowledge distillation objective is

$$\mathcal{L}_{\text{VKD}} = D_{\text{KL}}\left(\mathbf{p}_V \,\|\, \mathbf{p}_\theta^0\right). \tag{3}$$

$\mathcal{L}_{\text{VKD}}$ transfers the visual detector's relative confidence to the MLLM's initial-token distribution. During RBI, this soft target distinguishes confident and ambiguous detector outputs rather than treating them as identical hard labels. The resulting initial-token policy is then used to generate the student's on-policy rollouts in CT-OPSD.

Regularized classification for final decision. RBI trains the reflective response with standard autoregressive supervision:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{\sum_t m_t} \sum_t m_t \log p_\theta\left(y_t \mid x, q_{\text{eval}}, y_{<t}\right), \tag{4}$$

where $m_t \in \{0, 1\}$ is a binary loss mask. The real/fake class tokens in the `Initial Visual Judgment` and `Final Decision` are excluded by setting $m_t = 0$, while $m_t = 1$ for the remaining supervised tokens.

At the `Final Decision` token position, the MLLM `real`/`fake` logits $\mathbf{z}_\theta^1$ are obtained in the same manner. The MLLM and visual-detector logits are scaled using $T_{\text{LM}} = 3.0$ and $T_{\text{V}} = 1.0$, respectively, and then averaged to obtain $\mathbf{z}_{\text{avg}}$. The final classification objective is

$$\mathcal{L}_{\text{CLS}} = \text{CE}\left(\mathbf{z}_{\text{avg}}, c\right), \tag{5}$$

where c denotes the ground-truth label.

During RBI, $\mathcal{L}_{\text{CLS}}$ uses the visual-detector logits as a training-time regularizer for the final decision. For fixed MLLM logits, a smaller detector ground-truth margin, $z_V^c - z_V^{\bar{c}}$, lowers the averaged ground-truth margin. This increases the classification loss and the gradient applied to the MLLM logits relative to the case in which the detector confidently supports the correct class. A negative detector margin therefore encourages the MLLM to produce a larger correct-class margin to compensate for an incorrect initial judgment, whereas a large positive margin results in a smaller update. In this way, $\mathcal{L}_{\text{CLS}}$ supports keep-or-revise behavior according to the detector evidence. Logit averaging is used only during RBI training.

The overall RBI objective is

$$\mathcal{L}_{\text{RBI}} = \mathcal{L}_{\text{VKD}} + \mathcal{L}_{\text{SFT}} + \mathcal{L}_{\text{CLS}}. \tag{6}$$

Together, these objectives transfer the detector's confidence to the initial visual judgment, supervise the intervening reflective reasoning, and encourage the final decision to increase the ground-truth margin when the detector evidence is weak or incorrect.

3.3 CONTRASTIVE TEACHER ON-POLICY SELF-DISTILLATION

The first training stage internalizes the reflective decision pattern introduced in Equation 1. CT-OPSD further refines this behavior on reasoning trajectories generated by the student's current policy. To this end, we build on On-Policy Self-Distillation (OPSD) (Zhao et al., 2026; Shenfeld et al., 2026), which distills student-generated rollouts toward a privileged teacher at the token level.

Background: On-policy self-distillation. Given an image x and the evaluation prompt q_{eval}, the student samples an on-policy reflective rollout $\tau = (\tau_1, \ldots, \tau_L)$ from its current policy π_θ^S without access to privileged information. At position t, we define the student and privileged-teacher distributions as

$$S_t = \pi_\theta^S\left(\cdot \mid x, q_{\text{eval}}, \tau_{<t}\right), \qquad P_t = \pi_\phi^T\left(\cdot \mid x, q_{\text{eval}}, u, \tau_{<t}\right), \tag{7}$$

where u denotes privileged information available only to the teacher. The teacher observes the same student-generated prefix $\tau_{<t}$ while receiving additional information unavailable to the student, providing token-level guidance along the student's own rollout. Standard OPSD then optimizes

$$\mathcal{L}_{\text{OPSD}}(\theta) = \frac{1}{L}\sum_{t=1}^{L} D_{\text{KL}}\left(S_t \,\|\, P_t\right). \tag{8}$$

We introduce CT-OPSD, an extension of standard OPSD tailored to deepfake detection, which uses two contrasting privileged contexts to provide more targeted supervision for reflective reasoning.

Contrastive teacher token weighting. CT-OPSD constructs two privileged contexts for the same teacher model. Both contexts receive the visual-detector prediction $\hat{c}_V$, while one is additionally conditioned on the ground-truth label c and the other on its opposite $\bar{c}$. Thus, both contexts share the same initial visual judgment but guide reflection toward opposing final decisions. The same student rollout τ is teacher-forced under both contexts. At position t, the resulting teacher distributions are defined as

$$P_t^+ = \pi_\phi^T\left(\cdot \mid x, q_{\text{eval}}, \hat{c}_V, c, \tau_{<t}\right), \qquad P_t^- = \pi_\phi^T\left(\cdot \mid x, q_{\text{eval}}, \hat{c}_V, \bar{c}, \tau_{<t}\right), \tag{9}$$

where P_t^+ and P_t^- denote the teacher distributions conditioned on ground-truth and opposite labels, respectively. For the student-generated token τ_t, their disagreement is measured as

$$d_t = \left|P_t^+(\tau_t) - P_t^-(\tau_t)\right|. \tag{10}$$

The disagreement score d_t determines the relative contribution of each reasoning position to distillation. For positions $\mathcal{T}(\tau)$ inside the `think` section, the contrastive weight is defined as

$$w_t^{\text{CT}} = \frac{1 + \alpha d_t}{\frac{1}{|\mathcal{T}(\tau)|}\sum_{j\in\mathcal{T}(\tau)}(1+\alpha d_j)}, \qquad t \in \mathcal{T}(\tau), \tag{11}$$

with $w_t^{\text{CT}} = 1$ outside the `think` section. The additive constant prevents hard gating and ensures that every reasoning position retains a nonzero baseline contribution. The coefficient α controls the influence of teacher disagreement: $\alpha = 0$ recovers uniform weighting, equivalent to standard OPSD, while larger values place greater relative emphasis on positions with larger disagreement. The normalization keeps the average weight within the `think` section fixed at one. The effect of α is analyzed in Section C.1.

CT-OPSD objective. CT-OPSD distills the student toward the ground-truth-conditioned teacher distribution P_t^+, while P_t^- contributes only to the contrastive weight. The resulting objective is

$$\mathcal{L}_{\text{CT-OPSD}}(\theta) = \frac{1}{L}\sum_{t=1}^{L} a_t w_t^{\text{CT}} D_{\text{KL}}\left(S_t \,\|\, P_t^+\right), \tag{12}$$

where $a_t = 5$ for the real/fake class token in the `answer` section and $a_t = 1$ otherwise. The teacher policy π_ϕ^T is updated as an exponential moving average of the student throughout training. Consequently, reasoning positions with larger disagreement between the two privileged contexts have greater influence on the student update. Teacher disagreement is computed from full-vocabulary probabilities, whereas the reverse-KL term uses the student's renormalized top-20 support; further details are provided in Appendix A.4.

Table 1: **Cross-domain evaluation on DF40.** Models are evaluated on the unseen CDF domain, with recent manipulation methods reported for each forgery type. We report image-level balanced accuracy (%). Average results across all 32 manipulation methods are provided in Appendix C. Best and second-best results are shown in **bold** and underlined, respectively.

Method	FS			FR			EFS			FE	Avg.
	BlendFace	E4S	FaceDancer	SadTalker	MCNet	HyperReenact	PixArt-α	DiT-XL/2	SiT-XL/2	E4E	
Generic MLLMs											
LLaVA-1.5-7B	50.72	50.91	51.32	49.46	52.10	49.53	54.07	52.20	52.72	54.50	51.75
Qwen3-VL-8B-Instruct	50.21	53.76	51.43	50.27	50.56	50.13	66.64	49.90	50.19	53.31	52.64
InternVL3-8B-Instruct	48.95	58.67	52.70	47.58	45.14	49.65	58.05	52.23	52.58	35.16	50.07
MLLM-based Detectors											
SIDA-7B (Huang et al., 2025)	50.06	53.08	50.56	50.00	50.08	50.02	56.34	50.00	50.01	85.25	54.54
SIDA-13B (Huang et al., 2025)	50.11	74.96	63.08	49.90	51.51	50.24	75.78	50.28	50.45	**96.80**	61.31
X^2-DFD (Chen et al., 2026)	**69.63**	74.17	70.34	46.11	55.54	59.33	74.88	58.81	62.57	73.62	64.50
TranX-Adapter (Wang et al., 2026)	58.30	61.18	61.05	46.53	65.67	79.81	**93.45**	**74.28**	**78.83**	91.55	71.07
Deep-VRM (Lin et al., 2026)	49.69	62.11	50.37	53.16	48.62	50.55	84.65	59.65	60.63	84.36	60.38
Veritas (Tan et al., 2026b)	49.13	49.17	49.84	47.67	50.16	50.72	50.13	47.87	48.37	50.47	49.35
Veritas++ (Tan et al., 2026a)	49.94	66.81	52.39	49.92	51.70	51.35	68.90	52.10	52.18	70.49	56.58
SeeTR (ours)	65.47	**78.60**	**72.51**	**58.11**	**77.71**	**80.94**	78.13	69.64	70.78	79.30	**73.12**

Table 2: **Cross-domain & forgery evaluation on DF40.** Models are evaluated on datasets whose domains and manipulation methods are both unseen during training. We report image-level balanced accuracy (%). Best and second-best results are shown in **bold** and underlined, respectively.

Method	FS	FR	EFS	FE				Avg.
	DeepFaceLab	HeyGen	WhichIsReal	CollabDiff	StarGAN	StarGANv2	StyleCLIP	
Generic MLLMs								
LLaVA-1.5-7B	51.08	53.87	48.00	49.75	53.78	48.45	51.70	50.95
Qwen3-VL-8B-Instruct	51.57	70.15	53.05	50.80	60.86	50.38	60.17	56.71
InternVL3-8B-Instruct	52.62	69.83	56.74	29.10	63.03	49.72	26.12	49.60
MLLM-based Detectors								
SIDA-7B (Huang et al., 2025)	50.10	48.05	50.25	56.80	65.73	54.35	83.36	58.38
SIDA-13B (Huang et al., 2025)	52.69	64.65	50.75	75.75	76.46	59.65	92.89	67.55
X^2-DFD (Chen et al., 2026)	67.46	62.30	52.65	51.90	60.13	57.50	51.43	57.62
TranX-Adapter (Wang et al., 2026)	43.35	45.27	75.37	83.45	72.15	61.05	84.92	66.51
Deep-VRM (Lin et al., 2026)	51.30	76.46	**79.95**	73.10	55.42	50.58	**96.93**	69.11
Veritas (Tan et al., 2026b)	48.38	76.43	79.70	65.65	62.68	53.52	67.53	64.84
Veritas++ (Tan et al., 2026a)	55.82	77.97	78.82	81.15	72.78	54.02	88.72	72.76
SeeTR (ours)	**72.96**	**87.34**	67.33	**85.30**	**80.85**	**71.40**	75.08	**77.18**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Training details. We use InternVL3-8B-Instruct as the base MLLM and construct the training data from the FaceForensics++ (Rossler et al., 2019) domain of DF40 (Yan et al., 2024). To reduce shortcuts from resolution and compression artifacts (Grommelt et al., 2024), all images used to train and evaluate SeeTR are resized to 256×256 and encoded as JPEG with a quality factor of 90. Training uses a 12K class-balanced split shared by the visual detector and RBI, followed by a disjoint 5K split for CT-OPSD, sampled to maintain an approximate 8:2 ratio of correct and incorrect visual-detector judgments. Further training details are provided in Appendix A.2.

Evaluation protocol. Following the DF40 evaluation protocol (Yan et al., 2024), we evaluate generalization under two settings. The cross-domain setting evaluates models on the CDF (Li et al., 2020) domain after training on FF++, measuring generalization across data domains. The cross-domain & forgery setting uses datasets whose domains and manipulation methods are both unseen during training. All evaluations are performed at the image level, and baseline methods use their official code and default preprocessing pipelines.

At inference, the visual detector predicts `real` or `fake`, which is provided in the `<first>` field as the initial visual judgment. The MLLM then performs reflective reasoning conditioned on this judgment and produces the final decision in `<answer>`. This inference procedure follows the reflective behavior established during RBI, where the model learns to confirm or revise the visually grounded initial judgment. Further analysis is provided in Appendix C.1.

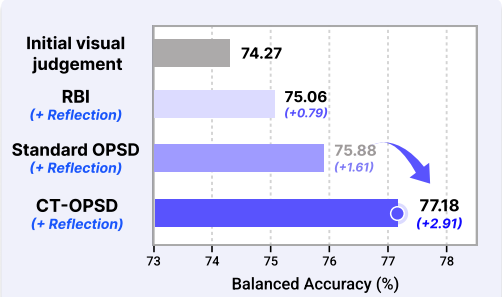

Figure 3: **Effect of CT-OPSD.** BAcc (%) under cross-domain & forgery evaluation. Standard OPSD improves upon RBI, while CT-OPSD further improves performance.

Table 3: **Effect of visual supervision on reflective behavior.** Balanced accuracy (%) under cross-domain & forgery evaluation. Starting from the same initial visual judgment, plain SFT degrades performance, whereas RBI improves it and enables further gains from CT-OPSD.

Stage 1 training	Before reflection	Reflection		BAcc Δ
		After Stage 1	After CT-OPSD	
Plain SFT	74.27	65.56 (−8.71)	63.82 (−1.74)	−10.45
RBI	74.27	**75.06** (+0.79)	**77.18** (+2.12)	+2.91

We use balanced accuracy (BAcc) as the primary metric because MLLM-based deepfake detectors can exhibit substantial class imbalance, as observed in Figure 1. BAcc assigns equal importance to both classes and is computed as $\mathrm{BAcc} = (\mathrm{Acc} * \mathrm{real} + \mathrm{Acc} * \mathrm{fake})/2$.

4.2 Experimental Results

Cross-domain evaluation. Table 1 evaluates generalization from the FF++ training domain to the unseen Celeb-DF (CDF) domain (Li et al., 2020). Averaged across the ten reported manipulation methods, SeeTR achieves a BAcc of **73.12%**, outperforming the strongest prior MLLM-based detector, TranX-Adapter, by **2.05** points. Notably, several recent MLLM-based detectors remain near 50–60% BAcc. The class-wise results in Table 8 reveal that this low BAcc often stems from **severe class skew**: SIDA and Veritas++ strongly favor real predictions, whereas Veritas exhibits the opposite bias. This mirrors the class-biased behavior identified in Figure 1, showing that the issue persists in recent MLLM-based detectors under domain shift. In contrast, SeeTR avoids such extreme class collapse and achieves the highest BAcc among the evaluated MLLM-based detectors.

Cross-domain & forgery evaluation. Table 2 evaluates generalization when both the real-source domain and forgery method are unseen during training. Following DF40 (Yan et al., 2024), we evaluate seven unseen forgery methods across four unseen real-source domains.[2] Averaged equally across the seven forgery methods, SeeTR establishes new state-of-the-art performance among the evaluated MLLM-based detectors, achieving a BAcc of **77.18%** and outperforming the strongest prior method, Veritas++, by **4.42** percentage points. SeeTR achieves the best performance on five methods: the FS and FR methods and three of the four FE methods.

4.3 Ablation Study

Effect of visual supervision on reflective behavior. Visual supervision enables RBI to internalize reflective behavior during Stage 1. To isolate its effect, we compare RBI with SFT using the same reflective responses and training data. Plain SFT applies autoregressive cross-entropy supervision to the entire response, including the real/fake tokens in both `<first>` and `<answer>`, whereas RBI follows the training objective defined in Equation 6. All other training settings are kept identical.

As shown in Table 3, both approaches are evaluated from the same initial visual judgment of 74.27% BAcc. After Stage 1, plain SFT reduces performance to 65.56%, whereas RBI improves it to 75.06%. This comparison shows that standard full-response SFT does not effectively preserve and use the initial visual judgment, while the visual supervision introduced in RBI better grounds the subsequent reflective decision process.

RBI also provides a stronger basis for subsequent CT-OPSD refinement. Applying CT-OPSD further reduces the plain-SFT result to 63.82%, whereas it improves RBI from 75.06% to 77.18%. Since

[2]MidJourney is included in the training data used by Veritas and Veritas++ (Tan et al., 2026b;a). To ensure a consistent evaluation protocol, we also include it in training and exclude it from evaluation.

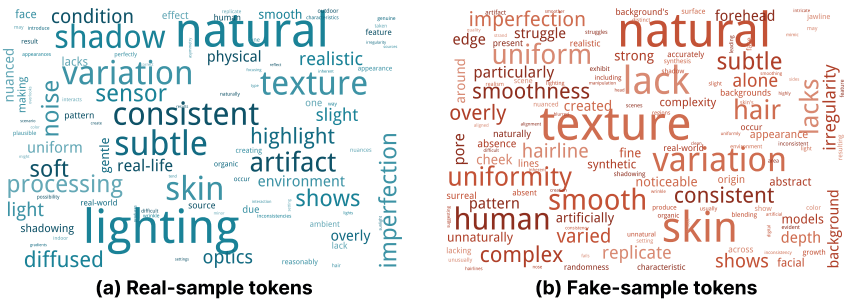

Figure 4: **Token-level disagreement in CT-OPSD.** Word clouds show tokens from reflective responses to 50 randomly sampled, correctly classified images per class: (a) real and (b) fake. Token size reflects the disagreement between the positive and negative teacher distributions. This qualitatively suggests that CT-OPSD places greater weight on reasoning positions associated with informative forensic cues.

Table 4: **Effect of reflection across unseen real-source domains.** Balanced accuracy (BAcc, %) is computed for each manipulation and then averaged within each domain. **Bold** indicates the better result before and after reflection for each domain.

SeeTR	CDF	UADFV	VFHQ	FFHQ	CelebA	Avg.
Before reflection	**74.10**	68.78	72.40	66.98	77.94	72.04
After reflection	73.12	**72.96**	**87.34**	**67.33**	**78.16**	**75.78**

both approaches start from the same detector-provided initial judgment at evaluation, the difference arises from how Stage 1 trains the model to use that judgment during reflection rather than from the initial judgment itself. These results indicate that effective on-policy refinement depends on first establishing the reflective behavior through RBI.

Effect of contrastive teacher weighting. Figure 3 shows that RBI improves BAcc from 74.27% to 75.06%, while standard OPSD further raises it to 75.88%. CT-OPSD achieves 77.18%, outperforming standard OPSD by 1.30 points and demonstrating the benefit of contrastive teacher weighting. Figure 4 further shows that large teacher disagreement is concentrated on tokens associated with informative forensic cues, supporting CT-OPSD's emphasis on such reasoning positions.

Effect of reflection across unseen real-source domains. We evaluate SeeTR across five unseen real-source domains in DF40: CDF (Li et al., 2020), UADFV (Li & Lyu, 2018), VFHQ (Xie et al., 2022), FFHQ (Karras et al., 2019), and CelebA (Liu et al., 2018). Table 4 shows that reflection improves four of the five domains, raising the domain-average BAcc from 72.04% to 75.78% (+3.74 points). This result demonstrates the effectiveness of our reflective reasoning in refining the initial visual judgment. Moreover, Table 9 shows that SeeTR achieves the highest domain-average BAcc among the MLLM-based detectors, establishing state-of-the-art performance under this setting.

5 Conclusion

In this work, we reveal that MLLM-based deepfake detection does not fully exploit the detection capability available in intermediate representations. Our analyses find higher probing performance at intermediate than later LM layers and class-biased predictions after reasoning-based SFT. We further observe that severe class skew persists in recent MLLM-based detectors, explaining their low BAcc under domain shifts. To address these limitations, we introduce SeeTR, which grounds an initial visual judgment in the most discriminative intermediate representation and explicitly trains subsequent reflection. RBI establishes effective reflective behavior where plain SFT degrades performance, while CT-OPSD further improves it over standard OPSD through contrastive teacher supervision. As a result, SeeTR achieves the highest BAcc among the evaluated MLLM-based detectors in both cross-domain and cross-domain & forgery settings, while reflection improves four of five unseen real-source domains. Together, these results demonstrate the effectiveness of explicitly grounding and refining visual judgments through reflective reasoning.

AI USE STATEMENT

We used Gemini-2.5-Flash to extract image-specific forensic artifact annotations for constructing RBI training data. The target MLLM then generated the reflective responses used for RBI; the external model did not generate these responses. Generative AI tools also assisted with code debugging, language editing, figure and table preparation, and discussion of experimental analysis. The authors reviewed the AI-assisted materials, verified the reported results, made all methodological decisions, and take responsibility for the content and conclusions of this work.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the visual detector, RBI and CT-OPSD objectives, inference procedure, and evaluation protocol. Appendix A provides the training data settings, hyperparameters, and implementation details for our method. Appendix B details the data construction and prompt design; Figures 5–10 illustrate this process and provide the prompts used for data generation, training, and evaluation.

ETHICS STATEMENT

We study deepfake detection using existing datasets of real and manipulated facial images. Detection errors and inaccurate forensic rationales could lead to mistaken conclusions about an image, particularly under domain shifts. Model outputs should therefore serve as supporting evidence with human review rather than the sole basis for consequential decisions. Descriptions of forensic artifacts could also inform attempts to evade detection; continued evaluation against new and adaptive manipulations remains important.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. Opd-v: Visual on-policy self-distillation with modality balance, 2026. URL https://arxiv.org/abs/2608.05131.

Rong Bao, Donglei Yu, Kai Fan, and Minpeng Liao. Fixing distribution shifts of llm self-critique via on-policy self-play training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17680–17700, 2025.

Jon Bateman. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace., 2022.

Lokendra Birla, Trishna Saikia, and Puneet Gupta. Avenue: A novel deepfake detection method based on temporal convolutional network and rppg information. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–16, 2025.

Matyas Bohacek and Hany Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48):e2216035119, 2022.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4103–4112. IEEE, 2022.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. *Advances in Neural Information Processing Systems*, 38:83792–83839, 2026.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1800–1807. Ieee, 2017.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes De Vreese. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1):69–91, 2021.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or jpeg? revealing common biases in generated image detection datasets, 2024. URL https://arxiv.org/abs/2403.17608.

Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 735–743, 2022a.

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

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

Inho Jung, Hyeongjun Choi, Minh Binh Le, Hohyun Na, and Simon Woo. A rich knowledge space for scalable deepfake detection. In *International Conference on Learning Representations*, volume 2026, pp. 96506–96535, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Pooja Kaushik, Vikas Garg, Anu Priya, and Shashi Kant. Financial fraud and manipulation: The malicious use of deepfakes in business. In *Deepfakes and their impact on business*, pp. 173–196. IGI Global Scientific Publishing, 2025.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning, 2026. URL https://arxiv.org/abs/2605.13255.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

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

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, and Shouhong Ding. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models, 2026. URL https://arxiv.org/abs/2606.15880.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 772–781. IEEE, 2021.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026. URL https://arxiv.org/abs/2605.21606.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16312–16321. IEEE, 2021.

Ethan Mendes, Jungsoo Park, and Alan Ritter. Making expert reasoning learnable with self-distillation. *arXiv preprint arXiv:2602.02405*, 2026.

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

Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 58(11):1–41, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL https://arxiv.org/abs/2601.19897.

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

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Zijian Yu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas++: Value-aware on-policy distillation for perception-enhanced aigi detection. *arXiv preprint arXiv:2607.27113*, 2026a.

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

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation, 2026. URL https://arxiv.org/abs/2606.05718.

Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6 (1):2056305120903408, 2020.

Wenbin Wang, Yuge Huang, Jianqing Xu, Yue Yu, Jiangtao Yan, Shouhong Ding, Pan Zhou, and Yong Luo. Tranx-adapter: Bridging artifacts and semantics within mllms for robust ai-generated image detection. *arXiv preprint arXiv:2602.21716*, 2026.

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

Simon Woo et al. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 122–130, 2022.

Jiahui Wu, Yu Zhu, Xiaoben Jiang, Yatong Liu, and Jiajun Lin. Local attention and long-distance interaction of rppg for deepfake detection: J. wu et al. *The Visual Computer*, 40(2):1083–1094, 2024.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 656–665. IEEE, 2022.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22355–22366. IEEE, 2023.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection, 2025a. URL https://arxiv.org/abs/2411.15633.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of clip in generalizable deepfake detection, 2025. URL https://arxiv.org/abs/2503.19683.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation, 2026. URL https://arxiv.org/abs/2605.18740.

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

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15024–15034. IEEE, 2021.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

APPENDIX

LIMITATIONS AND BROADER IMPACT

Limitations. Constructing RBI training data relies on reference images to extract artifact annotations for manipulated samples. We compare FS and FR images with their paired real sources, and EFS and FE images with fake references from the same generator or manipulation method. Such references may be unavailable for newly collected in-the-wild training data, limiting the extension of this annotation procedure. Reference images are not required for CT-OPSD training or inference. Obtaining suitable references for new data is a possible direction, but we have not evaluated it.

The visual detector adds a separate forward pass before reflective generation. Detector training updates only a linear classifier, but extracting its intermediate representation requires running the frozen base MLLM. At inference, the detector obtains the initial visual judgment from a frozen base MLLM, and the LoRA-adapted model generates the reflective response in a separate pass. Thus, the detector adds computation even though both models are based on the same pretrained backbone. Caching and sharing compatible vision-encoder features could reduce repeated visual encoding, although we have not implemented or evaluated this optimization.

Broader impact. This work advances deepfake detection by combining visual judgment with reflective forensic reasoning. By providing authenticity decisions alongside forensic rationales, SeeTR could support digital forensics, media verification, and efforts to identify manipulated content. Insights into detectable artifacts could also help adversaries conceal them, although our method is designed for detection rather than media generation. We hope this work contributes to more transparent assessment of online visual content.

A EXPERIMENTAL SETUP

A.1 VISUAL REPRESENTATIONS IN MLLMS AND EFFECT OF REASONING

Both experiments in Figure 1 use the FF++ training data of DF40 released by Veritas (Tan et al., 2026b; Yan et al., 2024). The evaluation set comprises ten of DF40's 32 manipulation methods in the unseen CDF domain, using the most recent methods from each forgery type: three FS, three FR, three EFS, and one FE, as listed in Table 1. We report image-level balanced accuracy averaged equally across the ten methods.

Table 5: **Optimization settings for the reasoning-effect experiment in Figure 1(b).**

Configuration	SFT	Linear probing
LoRA $(r, \alpha, \text{dropout})$	$(128, 256, 0.05)$	–
Learning rate	5×10^{-5}	5×10^{-4}
Optimizer	AdamW	AdamW
Batch size	16	16

Visual representations in MLLMs. For Figure 1(a), each MLLM receives an image together with the predefined evaluation prompt q_{eval}. We keep the MLLM frozen, mean-pool the visual-token hidden states from each selected language-model layer, and train a separate linear classifier using the DF40 training data. Since the evaluated MLLMs have different depths, we uniformly sample nine positions spanning their early, middle, and late layers. The indices 1–9 therefore indicate the relative order of the sampled layers rather than their original layer indices. Each layer-specific classifier is evaluated using the protocol described above.

Effect of reasoning. For Figure 1(b), we compare SFT with reasoning and visual-token linear probing using InternVL3-8B-Instruct. The SFT model is trained on the forensic rationales provided in the Veritas training set, whereas the linear classifier is trained on visual-token representations from the same DF40 images while the entire MLLM remains frozen. For linear probing, we use the ninth transformer block, which achieves the highest balanced accuracy on the validation set.

We report real accuracy, fake accuracy, and balanced accuracy under the same evaluation protocol. Table 5 summarizes the optimization settings.

A.2 Training details of SeeTR

Data splits. From the FF++ domain of DF40 (Yan et al., 2024), we draw 11,948 class-balanced images to train the visual detector. We use its predictions on these images as initial visual judgments to construct the RBI training set. We then apply the detector to the remaining FF++ images and sample 5,156 disjoint images for CT-OPSD training, with approximately 80% correct and 20% incorrect detector judgments. Training and evaluation are conducted at the image level. Table 6 summarizes the optimization settings for RBI and CT-OPSD. Both stages use InternVL3-8B-Instruct, updating only LM LoRA adapters while keeping the vision encoder and visual detector frozen.

Table 6: **Training configurations of RBI and CT-OPSD.**

Hyperparameter	RBI	CT-OPSD
Base model	InternVL3-8B-Instruct	InternVL3-8B-Instruct
Training samples	11,948	5,156
Training epochs	4	4
Effective batch size	16	32
Optimizer	AdamW	AdamW
Learning rate	5×10^{-5}	5×10^{-6}
Weight decay	0.01	0.01
Maximum gradient norm	1.0	0.1
LM LoRA rank / alpha / dropout	128 / 256 / 0.05	128 / 256 / 0.05
Teacher update	—	EMA, decay 0.995
KL vocabulary support	—	Student top-20 tokens

A.3 Details of vision detector

The vision detector consists of a frozen copy of the base MLLM and a linear classifier g. Given an image x and the evaluation prompt q_{eval}, we extract the hidden states corresponding to the visual `<IMG_CONTEXT>` tokens from the selected intermediate LM layer and mean-pool them to obtain $h_V(x, q_{\text{eval}})$.

Following the layer-wise probing analysis in Figure 1, we select the layer that achieves the highest balanced accuracy on a held-out validation split from the FF++ training domain. For InternVL3-8B-Instruct (Zhu et al., 2025), the best-performing representation is obtained from the output of the ninth language-model transformer block (block index 8 under zero-based indexing). Accordingly, we use `hidden_states[9]`, since `hidden_states[0]` corresponds to the embedding output.

We train only g on the pooled representation while keeping the base MLLM frozen. After training, both the MLLM and g remain fixed. The detector produces the visual logits $\mathbf{z}_V(x, q_{\text{eval}})$, and the higher-scoring class defines the initial visual judgment $\hat{c}_0$. These outputs provide the visual supervision used in RBI and determine the initial judgment for subsequent reflective refinement.

A.4 Details of CT-OPSD

Rollout and teacher update. CT-OPSD initializes both the student and teacher from the RBI-trained model. During on-policy rollout generation, the student receives only the image x and q_{eval} and generates the complete reflective response from its current policy, without access to the ground-truth label or the visual-detector prediction.

The generated rollout is then teacher-forced through two privileged contexts of the same teacher model. Both contexts include the visual-detector prediction $\hat{c}_V$. The positive context additionally contains the ground-truth label c, whereas the negative context contains its opposite $\bar{c}$. The teacher is maintained as an exponential moving average of the student and is treated as a stop-gradient model during distillation. The positive and negative teachers therefore correspond to two contexts of the same EMA teacher, rather than two independently trained models.

Contrastive weighting and distillation. For each student-generated token τ_t, the contrastive score s_t is computed from the absolute difference between the probabilities assigned by the positive and negative teachers. These probabilities are obtained from the full-vocabulary teacher distributions before any vocabulary truncation or renormalization.

The disagreement score contributes only to the contrastive weight w_t^{CT}, while reverse-KL distillation is performed against the positive teacher distribution. Specifically, the vocabulary support at each position is restricted to the 20 tokens with the highest student probabilities. The student and positive-teacher distributions are then renormalized over this shared support before computing the reverse KL. This top-20 restriction applies only to the distillation term and does not affect the full-vocabulary probabilities used for the contrastive score. No additional SFT or classification objective is applied during CT-OPSD. The remaining optimization and rollout settings are summarized in Table 6.

B Policy-Aligned Data Construction and Prompt Design

This section describes how we extract taxonomy-guided forensic evidence, generate reflective responses for RBI training, and construct label-conditioned teacher prompts for CT-OPSD.

Forensic artifact extraction. To provide image-specific forensic evidence for constructing reflective responses, we extract artifact annotations using Gemini-2.5-Flash. Its role is restricted to artifact localization and categorization rather than reasoning-response generation. The extraction is guided by the taxonomy in Table 7, which we curate by reviewing prior studies on deepfake artifacts and forgery categories (Pei et al., 2026; Yan et al., 2024) and manually inspecting samples across different forgery types. Each annotation specifies the artifact location, taxonomy category, visual cue, and a concise description.

For FS and FR, where corresponding real-source images are available, each manipulated image is compared with its real source, as illustrated in Figure 5. This comparison separates manipulation-induced defects from subject-specific facial characteristics and natural appearance changes. Gemini-2.5-Flash retains only visually distinct, taxonomy-consistent discrepancies and annotates up to three decisive artifacts in the manipulated image.

For EFS and FE, a corresponding real source is unavailable or not well defined. Each target image is therefore paired with another fake image produced by the same generator or manipulation method, as shown in Figure 6. Gemini-2.5-Flash first identifies an artifact shared by both images and then extracts additional image-specific artifacts, prioritizing distinct and unambiguous cues. Both extraction prompts reject subtle, uncertain, or taxonomy-inconsistent candidates. The resulting annotations are used only to construct reflective responses for manipulated images in the subsequent RBI stage.

Dataset construction for RBI. RBI responses are generated using two ground-truth-conditioned variants of the evaluation prompt q_{eval}. For each image, the visual-detector prediction $\hat{c}_V$ is inserted as the initial visual judgment $\hat{c}_0$. Real images use a prompt conditioned on $c = \mathrm{real}$, whereas manipulated images use an artifact-guided prompt conditioned on $c = \mathrm{fake}$ and the extracted artifact annotations (Figures 8 and 10, respectively).

The relation between $\hat{c}_V$ and c determines the target behavior: the response keeps the initial judgment when $\hat{c}_V = c$ and revises it when $\hat{c}_V \neq c$. The target MLLM then generates a reflective response y_i in the format of Equation 1. To keep these responses close to inference-time behavior, both prompts preserve the core task instruction, reasoning structure, and output constraints of q_{eval}. They introduce limited label-specific semantic guidance and rules that maintain a consistent reflective trajectory across Steps 1–3.

For RBI training, each generated response y_i is paired with the corresponding image x_i and the non-privileged evaluation prompt q_{eval} in Figure 7, yielding the RBI dataset defined in Equation 2. The ground-truth label and artifact annotations used for response generation are excluded from the RBI input. The same q_{eval} is used for visual-detector training and inference, RBI training, CT-OPSD student rollouts, and MLLM inference. Thus, the target MLLM generates the reasoning supervision, while privileged information is used only to guide its construction.

Prompt construction for CT-OPSD. After RBI establishes the reflective response behavior, CT-OPSD further refines it on the model's own rollouts. The student first generates an on-policy reflective rollout under q_{eval} without privileged information. The same fixed rollout is then evaluated by the teacher under real- and fake-conditioned contexts. The real-conditioned prompt in Figure 8 is reused as the real teacher context, while the label-conditioned prompt in Figure 9 serves as the fake teacher context.

Unlike the artifact-guided fake prompt used for RBI, the fake teacher prompt excludes image-specific artifact annotations. Both teacher contexts share the initial visual judgment and core response structure, while differing in their designated authenticity labels and limited label-specific semantic guidance. For each rollout, the positive context is conditioned on the ground-truth label c, whereas the negative context is conditioned on its opposite. Their token-level disagreement identifies reasoning positions where the teacher's guidance differs between the two contexts. CT-OPSD emphasizes the positive teacher's supervision at these positions, further refining the model's ability to confirm or revise its initial visual judgment.

Table 7: **Forgery-type artifact taxonomy for forensic evidence extraction.** The taxonomy comprises 59 artifact patterns across four forgery types: Face Swapping (FS), Face Reenactment (FR), Entire Face Synthesis (EFS), and Facial Editing (FE).

FS: Face Swapping	FR: Face Reenactment
Blending & boundary discontinuity • Boundary irregularity • Tone & color mismatch • Halo effect	**Structural misalignment** • Neck-head inconsistency • Iris distortion • Inconsistent shadow • Attribute loss
Spectrum & texture mismatch • Facial texture smoothing • Frequency & noise discrepancy • Compression & grid artifacts	**Boundary & masking anomalies** • Artificial seam line • Double edges / ghosting
Illumination & shadow inconsistency • Shadow inconsistency • Shadow omission	**Oral anatomical anomalies** • Dental structural malformation • Cavity texture blurring
Anatomical anomalies • Facial feature inconsistency • Facial feature misplacement • Gaze direction mismatch	**Expression disharmony** • Static upper face • Muscle activation mismatch
Segmentation & occlusion errors • Blended artifacts • Transparent objects	**Spectrum & texture mismatch** • Synthetic skin texture • Blurry lip boundary
	Spurious motion artifacts • Object / background distortion • Hair texture swimming

EFS: Entire Face Synthesis	FE: Facial Editing
Optical physics violations • Uniform sharpness • Illogical defocus • Resolution mismatch	**Structural & semantic inconsistency** • Object distortion • Background distortion • Anatomical disharmony
Material & surface texture bias • Digital airbrushing / wax-like skin • Painterly / muddy texture • Highlight clipping	**Texture & frequency degradation** • Over-smoothed texture • Noise & sharpness discrepancy
Structural asymmetry • Mismatched accessories • Irregular iris • Gaze deviation	**Residual & boundary artifacts** • Ghosting artifact • Texture fusion
Anatomical anomalies • Dental anomalies • Ear structure anomalies • Limb / hair anomalies	**Generative model-specific artifacts** • Material & surface texture bias • Contextual illumination mismatch • Logical hallucination
Contextual & spatial logic errors • Semantic pattern failure • Clothing pattern distortion • Hair & facial hair distortion • Background structure distortion • Spatial incoherence / blobs • Surreal / abstract background • Object hallucination	
Aesthetic & lighting bias • Hyper-realistic / volumetric lighting • Perfect symmetry • Compositional perfection	

C Additional Evaluation Results

Aggregate BAcc alone does not reveal how performance is distributed between real and fake images. We therefore report complete class-wise results for both evaluation settings in Tables 8 and 10. For SeeTR, both the initial visual judgment and final decision are included to show how reflection changes class-wise behavior.

Reflection across unseen real-source domains. Table 9 evaluates reflection across five unseen real-source domains: CDF (Li et al., 2020), UADFV (Li & Lyu, 2018), VFHQ (Xie et al., 2022), FFHQ (Karras et al., 2019), and CelebA (Liu et al., 2018). The initial visual judgment already achieves a strong domain-averaged BAcc of 72.04%, outperforming the prior MLLM-based detectors. Reflection further raises this performance to 75.78%, yielding a 3.74-point improvement. It improves six of the eight evaluation units, including gains of 14.94 points on HeyGen and 4.18 points on DFL. These results show that SeeTR builds effective reflection on a strong visual baseline across diverse real-source domains.

Cross-domain evaluation. Table 8 reports results across all 32 manipulation methods in the CDF domain. Reflection improves *Avg. Fake* from 76.47% to 83.65%, with gains across all four forgery types. This improvement is offset by lower accuracy on the shared CDF real set, resulting in a modest 0.98-point decrease in BAcc, from 74.10% to 73.12%. Nevertheless, the final decision achieves the highest BAcc among all competing methods.

Table 8: **Class-wise results for cross-domain evaluation.** We report fake accuracy (%) by forgery type across all 32 manipulation methods in the CDF domain. *Avg. Fake* is averaged over all 32 methods. Since all fake subsets share the same CDF real set, real accuracy is reported once per method. The corresponding BAcc results are reported in Table 9.

Method	FS Fake	FR Fake	EFS Fake	FE Fake	Avg. Fake	Real
Vision-based detectors						
Xception (Chollet, 2017)	60.95	61.97	60.75	79.76	61.86	59.63
RECCE (Cao et al., 2022)	57.91	52.26	54.82	66.63	55.10	71.59
UCF (Yan et al., 2023)	47.67	41.73	46.20	57.77	45.30	80.87
SPSL (Liu et al., 2021)	87.37	88.00	79.14	82.56	84.88	27.29
SRM (Luo et al., 2021)	70.18	71.82	74.48	77.20	72.36	56.82
Effort (Yan et al., 2025a)	90.33	60.46	93.53	99.64	80.42	64.70
Generic MLLMs						
LLaVA-1.5-7B (Liu et al., 2024)	10.19	9.45	14.56	16.17	11.47	92.83
Qwen3-VL-8B-Instruct (Yang et al., 2025)	3.18	1.10	6.07	1.63	3.26	98.82
InternVL3-8B-Instruct (Zhu et al., 2025)	41.10	28.54	47.97	6.32	37.45	63.99
MLLM-based detectors						
SIDA-7B (Huang et al., 2025)	1.04	0.09	3.11	70.50	3.50	100.00
SIDA-13B (Huang et al., 2025)	12.53	3.47	31.21	94.44	17.53	99.16
X^2-DFD (Chen et al., 2026)	92.72	67.42	93.47	97.44	83.61	49.79
TranX-Adapter (Wang et al., 2026)	31.71	39.85	91.35	96.19	55.42	86.92
Deep-VRM (Lin et al., 2026)	33.10	26.24	69.24	96.98	43.82	71.73
Veritas (Tan et al., 2026b)	92.59	95.97	94.67	98.41	94.69	2.53
Veritas++ (Tan et al., 2026a)	12.86	6.01	30.02	44.36	16.64	96.62
SeeTR (*Before reflection*)	73.35	75.06	75.82	92.06	76.47	71.73
SeeTR (*After reflection*)	81.80	81.92	83.10	96.02	83.65	62.59

Table 9: **BAcc (%) before and after reflection across unseen real-source domains.** Each meta-column groups evaluation sets by their real-image source domain. CDF aggregates all 32 manipulation methods. DFL and WIR denote DeepFaceLab and WhichIsReal, respectively. *Domain Avg.* first averages the evaluation sets within each source domain and then averages the five domain-level results. Arrows indicate dataset-level improvements after reflection. Best and second-best results are shown in **bold** and underlined, respectively.

Method	CDF	UADFV	VFHQ	FFHQ	CelebA				Domain Avg.
	DF40 (32 methods)	DFL	HeyGen	WIR	CollabDiff	StarGAN	StarGANv2	StyleCLIP	
Generic MLLMs									
LLaVA-1.5-7B (Liu et al., 2024)	52.15	51.08	53.87	48.00	49.75	53.78	48.45	51.70	51.20
Qwen3-VL-8B-Instruct (Yang et al., 2025)	51.04	51.57	70.15	53.05	50.80	60.86	50.38	60.17	56.27
InternVL3-8B-Instruct (Zhu et al., 2025)	50.72	52.62	69.83	56.74	29.10	63.03	49.72	26.12	54.38
MLLM-based Detectors									
SIDA-7B (Huang et al., 2025)	51.75	50.10	48.05	50.25	56.80	65.73	54.35	83.36	53.04
SIDA-13B (Huang et al., 2025)	58.34	52.69	64.65	50.75	75.75	76.46	59.65	92.89	60.52
X^2-DFD (Chen et al., 2026)	66.70	67.46	62.30	52.65	51.90	60.13	57.50	51.43	60.87
TranX-Adapter (Wang et al., 2026)	71.17	43.35	45.27	75.37	83.45	72.15	61.05	84.92	62.11
Deep-VRM (Lin et al., 2026)	57.77	51.30	76.46	**79.95**	73.10	55.42	50.58	**96.93**	66.90
Veritas (Tan et al., 2026b)	48.61	48.38	76.43	79.70	65.65	62.68	53.52	67.53	63.09
Veritas++ (Tan et al., 2026a)	56.63	55.82	77.97	78.82	81.15	72.78	54.02	88.72	68.68
SeeTR (*Before reflection*)	**74.10**	68.78	72.40	66.98	84.55	80.32	**73.45**	73.42	72.04
SeeTR (*After reflection*)	73.12	**72.96**	**87.34**	67.33	**85.30**	**80.85**	71.40	75.08	**75.78**

Table 10: **Class-wise results for cross-domain & forgery evaluation.** We report real and fake accuracy (%) for each evaluation set, together with their unweighted averages across the seven sets. Several MLLM-based detectors exhibit large real–fake accuracy gaps, revealing substantial class skew under domain and forgery shifts. DFL and WIR denote DeepFaceLab and WhichIsReal, respectively. The corresponding BAcc results are reported in Table 9.

Method	CollabDiff		DFL		HeyGen		StarGAN		StarGANv2		StyleCLIP		WIR		Avg.	
	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake
Vision-based detectors																
Xception	95.30	53.50	63.61	67.55	84.42	61.56	83.57	31.30	79.85	16.15	95.50	64.66	66.30	33.47	81.22	46.88
RECCE	94.00	40.20	63.09	54.20	75.52	40.81	88.36	20.92	91.05	4.85	94.90	46.85	59.90	37.36	80.97	35.03
UCF	94.00	29.10	74.14	58.36	83.09	40.37	89.31	33.37	89.30	9.20	94.50	39.24	84.10	11.79	86.92	31.63
SPSL	34.90	87.80	29.22	93.40	38.07	69.22	26.92	87.15	29.45	68.25	31.90	79.43	43.90	47.85	33.48	76.16
SRM	81.80	76.30	45.96	81.93	44.03	58.13	78.53	38.26	76.55	19.35	82.90	72.66	29.80	76.22	62.80	60.41
Effort	78.20	51.30	69.17	91.22	83.78	66.92	66.68	99.90	76.50	80.55	78.70	83.50	76.10	23.08	75.59	70.92
Generic MLLMs																
LLaVA-1.5-7B	89.20	10.30	95.41	6.76	96.10	11.65	89.36	18.20	90.59	6.31	88.10	15.30	90.80	5.19	91.37	10.53
Qwen3-VL-8B-Instruct	96.30	5.30	99.48	3.65	99.99	40.31	95.92	25.81	97.70	3.05	95.20	25.14	99.70	6.39	97.75	15.67
InternVL3-8B-Instruct	36.90	21.30	72.72	32.51	88.32	51.34	78.78	47.28	85.99	13.46	37.30	14.94	98.60	14.89	71.23	27.96
MLLM-based detectors																
SIDA-7B	99.90	13.70	100.00	0.19	92.79	3.30	91.78	39.67	96.75	11.95	99.90	66.82	99.90	0.60	97.29	19.46
SIDA-13B	98.10	53.40	100.00	5.38	97.47	31.84	55.29	97.63	74.60	44.70	97.40	88.39	99.60	1.90	88.92	46.18
X^2-DFD	5.80	98.00	36.59	98.33	24.85	99.75	20.51	99.75	16.30	98.70	5.30	97.56	6.40	98.90	16.54	98.71
TranX-Adapter	99.90	67.00	52.29	34.40	31.34	59.19	47.38	96.93	27.10	95.00	99.90	69.94	81.60	69.13	62.79	70.23
Deep-VRM	98.80	47.40	76.08	26.52	61.70	91.21	94.20	16.63	95.85	5.30	98.50	95.36	59.90	100.00	83.58	54.63
Veritas	42.60	88.70	6.08	90.68	55.41	97.45	35.64	89.72	16.25	90.80	40.80	94.26	60.80	98.60	36.80	92.89
Veritas++	98.10	64.20	98.38	13.26	93.64	62.31	97.43	48.14	92.85	15.20	97.90	79.53	93.30	64.34	95.94	49.57
SeeTR (*Before reflection*)	84.80	84.30	87.78	49.78	98.57	46.23	78.28	82.36	68.12	78.78	82.70	64.13	91.30	42.66	84.51	64.03
SeeTR (*After reflection*)	78.20	92.40	72.92	73.00	97.66	77.01	70.61	91.08	57.56	85.24	75.80	74.36	90.20	44.46	77.56	76.79

C.1 Additional Ablation Studies

Effect of detector-provided initial judgment. Our main evaluation uses the visual detector prediction as the initial visual judgment in <first> before reflective reasoning. Accordingly, the performance before reflection in this setting directly corresponds to the vision detector. To examine whether the performance of SeeTR depends on providing this detector prediction at inference, we additionally evaluate a full-generation setting in which the detector output is not provided and the MLLM autoregressively generates <first>, <think>, and <answer>. In this setting, the performance before reflection is measured from the MLLM-generated prediction in <first>.

As shown in Table 11, the MLLM-generated initial judgment achieves 74.10% BAcc, only 0.17 percentage points below the detector-provided initial judgment of 74.27%. Although RBI distills the visual-detector distribution into the MLLM's initial judgment, this supervision does not constrain the MLLM-generated <first> to exactly reproduce the detector prediction at inference. This difference leads to a small performance gap between the two initial judgments.

Importantly, the reflective behavior is preserved under full generation. Starting from the MLLM-generated initial judgment, reflection improves BAcc from 74.10% to 76.78%, corresponding to a gain of 2.68 percentage points. This is comparable to the 2.91-point gain obtained in the detector-provided setting, where performance improves from 74.27% to 77.18%. Thus, the reflection gain does not rely solely on directly providing the detector prediction at inference, and the model retains the learned reflective behavior when generating the complete response autoregressively. Since the detector-provided setting yields the higher final performance, we use it as the main inference protocol throughout our experiments.

Table 11: **Effect of detector-provided initial judgment.** We compare the main evaluation setting, where the vision detector provides the initial judgment in <first>, with full generation, where the MLLM generates <first> itself. Accordingly, "Before reflection" denotes the vision-detector prediction in the detector-provided setting and the MLLM-generated <first> prediction in the full-generation setting. "After reflection" denotes the final MLLM prediction in <answer>. We report image-level balanced accuracy (%) under the cross-domain & forgery setting.

Evaluation setting	FS	FR	EFS	FE				Avg.
	DeepFaceLab	HeyGen	WhichIsReal	CollabDiff	StarGAN	StarGANv2	StyleCLIP	
Full generation								
Before reflection	70.71	72.93	63.08	83.85	**82.99**	**72.82**	72.29	74.10
After reflection	**73.34**	**87.16**	**63.13**	**85.45**	81.53	71.72	**75.10**	**76.78**
Detector-provided initial judgment								
Before reflection	68.78	72.40	66.98	84.55	80.32	**73.45**	73.41	74.27
After reflection	**72.96**	**87.34**	**67.33**	**85.30**	**80.85**	71.40	**75.08**	**77.18**

Sensitivity to the disagreement coefficient α. We further examine the sensitivity of CT-OPSD to the disagreement coefficient α, which controls the relative emphasis placed on reasoning positions with larger positive–negative teacher disagreement. As shown in Table 12, CT-OPSD achieves 76.01%, 77.02%, 77.18%, and 77.00% BAcc with $\alpha = 1.0$, 1.5, 2.0, and 2.5, respectively. The best performance is obtained at $\alpha = 2.0$, which is used throughout our main experiments. Overall, CT-OPSD maintains comparable performance across the tested values of α, demonstrating robustness to the choice of the disagreement coefficient.

Table 12: **Sensitivity of CT-OPSD to the hyperparameter α under the cross-domain & forgery setting.** We report image-level BAcc (%) averaged over all manipulation methods.

α	BAcc	Δ
1.0	76.01	–
1.5	77.02	+1.00
2.0	**77.18**	**+1.17**
2.5	77.00	+0.99

Further analysis of contrastive teacher weighting. To understand where CT-OPSD improves over standard OPSD, we analyze reflection behavior conditioned on the visual detector's initial judgment. Table 13 partitions 29,599 samples from the cross-domain & forgery evaluation by the ground-truth label and whether the initial judgment is correct. Each entry is reported as n/N, where N denotes the number of samples in each initial-judgment case and n the number for which reflection takes the desired action. Accordingly, *Keep real* and *Keep fake* measure preservation of correct judgments, while *Flip fake→real* and *Flip real→fake* measure correction of initially wrong judgments.

Table 13: **Reflection behavior of standard OPSD and CT-OPSD.** Successful keep-or-flip cases are reported as n/N over 29,599 samples from the cross-domain & forgery evaluation.

Reflection behavior	Standard OPSD	CT-OPSD
GT: Real		
Keep real	$\mathbf{13{,}989}/14{,}288$	$13{,}456/14{,}288$
Flip fake → real	$2/1{,}777$	$\mathbf{25}/1{,}777$
GT: Fake		
Flip real → fake	$1{,}020/4{,}949$	$\mathbf{1{,}918}/4{,}949$
Keep fake	$\mathbf{8{,}585}/8{,}585$	$8{,}576/8{,}585$
Corrective flips	$1{,}022$	$1{,}943$
Corrective reflection	$+723$	$\mathbf{+1{,}102}$

The largest improvement of CT-OPSD appears for fake images initially judged real: it correctly flips 1,918 of 4,949 cases, compared with 1,020 for standard OPSD. In contrast, recovery of real images initially judged fake remains limited, with 25 of 1,777 cases corrected by CT-OPSD versus 2 by standard OPSD. Consistently, *Keep real* decreases from 13,989 to 13,456, while the total number of corrective flips increases from 1,022 to 1,943 and net corrective reflection rises from +723 to +1,102.

These results reveal an asymmetric reflection behavior: CT-OPSD is particularly effective at recovering unseen fake images missed by the initial visual judgment, but remains less effective at correcting decisions toward the real class. Despite this fake-oriented tendency, SeeTR consistently achieves the highest BAcc among the evaluated MLLM-based detectors under both cross-domain and cross-domain & forgery settings, as well as the highest domain-average BAcc across unseen real-source domains. This suggests that contrastive teacher weighting substantially strengthens corrective reflection for missed fake images, while leaving room to further balance reflection between the two classes.

Artifact extraction (original real source)



Image A
(Reference)



Image B
(Target)

```
# Fake
Role. You are an expert AI forensics analyst specializing in deepfake detection.

I will provide two images:
1. **Reference Image:** The **ABSOLUTE GROUND TRUTH**. This defines what is "natural" for this specific subject.
2. **Investigation Image:** The manipulated (fake) image generated using the **'{forgery_type}'** technique.

**Context:**
Face Swapping (FS)** is a technique for synthesizing the identity of a source person by transplanting it into the face area of a target image.

**[Knowledge Base: Forensic Artifact Taxonomy {forgery_type}]**
Classify strictly using these fake artifact taxonomy.

{fake_artifact_taxonomy}

**[CRITICAL INSTRUCTION: The Differential & Quality Protocol]**
Before generating any output, you must perform the following logical steps internally:

1.  **Step 1: Subtraction (Difference Check).** Compare the Investigation Image against the Reference Image. Identify what is new or different.
2.  **Step 2: Validation (The "Defect" Rules).** Validate the identified differences strictly against these three mandatory rules:
    * **Rule A (Natural Trait Check):** If the feature exists in the Reference Image, it is **NOT** an artifact. **IGNORE IT.**
    * **Rule B (Visual Realism Check):** Even if a feature is NEW (not in Reference), does it look visually natural?
        * If it is a realistic change (e.g., mouth open naturally, slight head turn that looks physically correct), **IGNORE IT.** We are looking for **FLAWS**, not just
changes.
    * **Rule C (Taxonomy Compliance Check):** Can you classify this error strictly into one of the 5 categories in the Knowledge Base?
        * If the anomaly cannot be clearly defined by the provided Visual Cues (e.g., "it just looks weird" but no specific flaw), **DO NOT REPORT IT.**
        * **Do not force a classification.** If it doesn't fit the rules, discard it.
3.  **Step 3: Description.** Describe ONLY the validated technical artifacts strictly as visual flaws within the Investigation Image.

**Instructions for Output:**
1.  **Strict Selection:** Identify the **most distinct technical artifacts (1 to 3 maximum)** that passed ALL Steps (1, 2, and 3).
    * **Zero Tolerance for Ambiguity:** If an artifact is subtle or debatable, do not report it.
    * **Do NOT force 3 artifacts:** If only 1 distinct technical flaw exists, report only 1.
    * **Analysis Rejection Option:** If you cannot find any valid, distinct artifacts that meet the strict criteria above, output exactly: **"No artifacts found."**
2.  **Format:** Use the structure below.

**[Output Structure & Example]**
*(Describe the visual flaw strictly. Do not mention "Reference Image" in the final text.)*
*(Description has only one sentence)*

**Artifact 1**
- **Where:** ...
- **Category:** ...
- **Visual Cue:** ...
- **Description:** ...
"""
```

Figure 5: **Taxonomy-guided artifact annotation with a paired real source.** For Face Swapping and Face Reenactment, the manipulated target image is compared with its corresponding real source image. Gemini-2.5-Flash first identifies differences introduced in the target and then removes candidates attributable to subject-specific characteristics, natural appearance changes, or visually plausible variations. The remaining candidates are validated against the forgery-type taxonomy, and only the most decisive and visually observable artifacts are retained. For each selected artifact, the prompt annotates its spatial location, taxonomy category, visual cue, and a concise description. Up to three artifacts are reported, while ambiguous or taxonomy-inconsistent candidates are discarded.

Artifact extraction (same manipulation technique)



Image A
(Reference)



Image B
(Target)

```
# Fake
You are an expert AI Forensics Analyst specializing in Deepfake Detection.

  I will provide two images:
  1. **Investigation Image A:** A synthetic face generated by a specific AI model.
  2. **Investigation Image B:** Another synthetic face generated by the **SAME** AI model.

  **Context:**
  Both images were created using the **'{forgery_type}'** technique (Entire Face Synthesis).
  EFS generates non-existent identities from random noise. Artifacts typically fall into two main spectrums depending on the model architecture:
  * **GAN-based (e.g., StyleGAN):** Excellent skin texture (pores visible) but prone to **structural asymmetry** (mismatched earrings/eyes) and **abstract/
incoherent backgrounds** (blobs).
  * **Diffusion-based (e.g., Midjourney, Flux):** High aesthetic quality (cinematic lighting) but prone to **physical logic failures** (blending objects, gravity defiance)
and **painterly/smoothed textures** in detailed areas.
  **Objective:**
  Perform a comparative forensic analysis to identify synthetic artifacts.
  1. Identify exactly **one "Shared Artifact"** that appears in **BOTH** images, considering the {forgery_type} (AI model specific artifacts).
  2. Analyze **Image A** individually to detect **up to 2 distinct artifacts** (excluding the Shared Artifact).
  3. Analyze **Image B** individually to detect **up to 2 distinct artifacts** (excluding the Shared Artifact).

**[Knowledge Base: Forensic Artifact Taxonomy {forgery_type}]**
Classify strictly using these fake artifact taxonomy.

{fake_artifact_taxonomy}


  **[CRITICAL INSTRUCTION: The Analysis Protocol]**
  Execute the following steps internally:
  1.  **Step 1: The "Focus & Logic" Filter (Anti-Hallucination).**
    * **Scan for Blur:** Identify areas that are out-of-focus due to depth-of-field (background, ears, shoulders, hair edges).
    * **STRICT PROHIBITION:** You are **strictly prohibited** from flagging an area as *"Contextual & Spatial Logic Errors"* or *"Anatomical Anomalies"* if it is
blurred or out-of-focus.
  2.  **Step 2: Commonality Check (Priority).** Scan both images to find an artifact category/cue that exists in **BOTH** (e.g., Wax-like Skin). Select this as the
**Shared Artifact**.
  3.  **Step 3: Individual Selection (Anti-Redundancy).**
    * Select max 2 unique artifacts for **Image A**.
    * Select max 2 unique artifacts for **Image B**.
    * **Constraint:** To maximize coverage, **PRIORITIZE artifacts that are DIFFERENT from the Shared Artifact.**
    * *Example:* If "Wax-like Skin" is the Shared Artifact, look for other errors for the Individual Image sections. Only list "Wax-Like Skin" again if it is the absolutely
ONLY flaw visible.
  4.  **Step 4: Drafting.** Write descriptions.
    * **Rule:** Even for the Shared Artifact, describe it as if observing a single instance of a flaw. Do NOT use comparison phrases like "Both images show...".

  **Instructions for Output:**
  1.  **Selection Strategy:**
    * **Shared Artifact:** 1 Mandatory Artifact (present in both).
    * **Image A/B:** Max 2 Artifacts **(Preferably distinct from the Shared Artifact)**.
  2.  **Zero Tolerance:** If an artifact is subtle or debatable, do not report it.
  3.  **Rejection:** If no valid artifacts are found in a section, output exactly: **"No artifacts found."**
  4.  **Format:** Use the exact structure below. Description has only one sentence.
  5.  **Visual Cue:** If you can not find Visual Cue in the knowledge base, but fits Category definition, write Category name (e.g., "Material & Surface Texture Bias")

  **[Output Structure & Example]**
  **[Analysis: Investigation Image A]**
  **Artifact 1**
  - **Where:** ...
  - **Category:** ...
  - **Visual Cue:** ...
  - **Description:** ...
```

Figure 6: **Taxonomy-guided artifact annotation without a paired real source.** For Entire Face Synthesis and Facial Editing, where a corresponding real source is unavailable or not well defined, two fake images produced by the same generator method are jointly analyzed. Gemini-2.5-Flash first identifies one shared artifact observed in both images, capturing a recurring trace of the underlying generation process. It then extracts up to two additional image-specific artifacts from each image while avoiding cues that are redundant with the shared artifact. Ambiguous, out-of-focus, or weakly supported candidates are rejected. Each retained artifact is annotated with its spatial location, taxonomy category, visual cue, and a concise description.

Evaluation prompt

SYSTEM_PROMPT:
<image>
You are a forensic image analyst. Determine whether this facial image is real or fake by following the instructions.",

USER_PROMPT:
Determine whether the facial image is a real camera-captured facial photo or an AI-generated facial image.

Output Structure
The response must follow a <first>–<think>–<answer> structure.
In <first>, write your initial visual judgment as either \"real\" or \"fake\" based only on your first visual impression.

Reasoning Process
Write the forensic reasoning inside <think> in exactly 3 steps.
Step 1. Initial judgment awareness: acknowledge that the visual first impression written in <first> is only an initial observation, and state that the final decision must be based on closer image-grounded evidence rather than that first impression alone.
Step 2. Forensic evidence search: do not simply follow the visual first impression in <first>; independently re-check the image and identify the single strongest image-grounded visual evidence relevant to the real-versus-fake decision. Clearly state where the evidence appears and what specific visual pattern is observed.
Step 3. Evidence interpretation: deeply interpret the evidence identified in Step 2. Compare whether the observed pattern is better explained by natural camera capture, physical lighting, optics, sensor noise, camera processing, compression, pose, anatomy, or ordinary physical image formation, or by AI generation. Explain whether this evidence supports or opposes the initial visual judgment in <first>, and determine the final decision.

Final Decision
Write the final verdict inside <answer> as exactly \"real\" or \"fake\".

STRICT RULES
Step 1 must be exactly one short sentence. Steps 2 and 3 must each contain 2 to 3 complete sentences. Steps 2 and 3 must use the same single visual evidence.

Figure 7: **Evaluation prompt.** This prompt contains no ground-truth label or forensic artifact information. It is provided with the image during visual-detector training and inference, and is used as the MLLM input for RBI training, CT-OPSD student rollouts, and inference. Thus, both components operate under the same non-privileged input context.

Real-conditioned response-generation prompt (for RBI & OPSD)

SYSTEM_PROMPT:
<image>
You are a forensic image analyst. Determine whether this facial image is real or fake by following the instructions.

USER_PROMPT:
Determine whether the facial image is a real camera-captured facial photo or an AI-generated facial image.

Output Structure
The response must follow a <first>–<think>–<answer> structure.
In <first>, write your initial visual judgment as either \"real\" or \"fake\" based only on your first visual impression.

Reasoning Process
Write the forensic reasoning inside <think> in exactly 3 steps.
***Step 1. Initial judgment awareness:** acknowledge that the visual first impression written in <first> is only an initial observation, and state that the final decision must be based on closer image-grounded evidence rather than that first impression alone.*
***Step 2. Forensic evidence search:** do not simply follow the visual first impression in <first>; independently re-check the image and identify the single strongest image-grounded visual evidence that supports a real camera-captured interpretation. Clearly state where the evidence appears and what specific visual pattern is observed.*
***Step 3. Evidence interpretation:** deeply interpret the evidence identified in Step 2 and explain why it is more consistent with a real camera-captured image than with AI generation. Explain why treating the observed pattern as a fake artifact would be an overclaim and how it can plausibly arise from physical lighting, optics, natural sensor noise, camera processing, compression, pose, anatomy, or other ordinary photographic phenomena. If the visual first impression in <first> suggested fake, explain why that impression overemphasized a superficial cue and why closer evidence supports real. If it suggested real, explain why closer evidence confirms it. Never mention that the target label was provided in advance.*

Final Decision
Write the final verdict inside <answer> as exactly \"real\".

STRICT RULES
*The visual first impression is \"{**first_prediction**}\", but the final answer should be \"real\"; use the first impression only as context, do not justify it when it conflicts with real, and write <think> so that the evidence supports <answer>real</answer>.*
Step 1 must be exactly one short sentence. Steps 2 and 3 must each contain 2 to 3 complete sentences.
Steps 2 and 3 must use the same single visual evidence.

Figure 8: **Real-conditioned response-generation prompt.** The red text marks the case-specific context inserted into the prompt, including the initial visual judgment and the real ground-truth label; the latter serves as privileged guidance for correct reflection. This prompt is used both to generate RBI responses for real images and as the real-conditioned teacher context in CT-OPSD.

Fake-guided response-generation prompt (for OPSD)

SYSTEM_PROMPT:
<image>
You are a forensic image analyst. Determine whether this facial image is real or fake by following the instructions.

USER_PROMPT
Determine whether the facial image is a real camera-captured facial photo or an AI-generated facial image.

Output Structure
The response must follow a <first>–<think>–<answer> structure.
In <first>, write your initial visual judgment as either \"real\" or \"fake\" based only on your first visual impression.

Reasoning Process
Write the forensic reasoning inside <think> in exactly 3 steps.
***Step 1. Initial judgment awareness:** acknowledge that the visual first impression written in <first> is only an initial observation, and state that the final decision must be based on closer image-grounded evidence rather than that first impression alone.*
***Step 2. Forensic evidence search:** do not simply follow the visual first impression in <first>; independently re-check the image and identify the single strongest image-grounded visual evidence supporting a fake interpretation. Clearly state where the evidence appears and what specific visual pattern is observed.*
***Step 3. Evidence interpretation:** deeply interpret the evidence identified in Step 2 and explain why it is more consistent with AI generation than with natural image formation. If the visual first impression in <first> suggested real, explain what it overlooked and why closer evidence supports fake. If it suggested fake, explain why closer evidence confirms it. Never mention that the target label was provided in advance.*

Final Decision
Write the final verdict inside <answer> as exactly \"fake\".

STRICT RULES
*The visual first impression is \"{**first_prediction**}\", but the final answer should be \"fake\"; use the first impression only as context, do not justify it when it conflicts with fake, and write <think> so that the evidence supports <answer>fake</answer>.*
Step 1 must be exactly one short sentence. Steps 2 and 3 must each contain 2 to 3 complete sentences. Steps 2 and 3 must use the same single visual evidence."

Figure 9: **Fake-conditioned CT-OPSD teacher prompt.** The red text marks the initial visual judgment and the fake ground-truth label provided to the teacher. This prompt excludes image-specific artifact information and is used as the fake-conditioned teacher context in CT-OPSD, keeping the privileged guidance comparable between the real- and fake-conditioned teacher contexts.

Fake & artifact -guided response-generation prompt (for RBI)

SYSTEM_PROMPT:
<image>
You are a forensic image analyst. Determine whether this facial image is real or fake by following the instructions.

USER_PROMPT
Determine whether the facial image is a real camera-captured facial photo or an AI-generated facial image.

Output Structure
The response must follow a <first>–<think>–<answer> structure.
In <first>, write your initial visual judgment as either \"real\" or \"fake\" based only on your first visual impression.

Reasoning Process
Write the forensic reasoning inside <think> in exactly 3 steps.
***Step 1. Initial judgment awareness:** acknowledge that the visual first impression written in <first> is only an initial observation, and state that the final decision must be based on closer image-grounded evidence rather than that first impression alone.*
***Step 2. Forensic evidence search:** do not simply follow the visual first impression in <first>; independently re-check the image, paying particular attention to \"{**artifact_region**}\", and identify the single strongest image-grounded visual evidence in that region. Clearly state where the evidence appears and what specific visual pattern is observed.*
***Step 3. Evidence interpretation:** deeply interpret the evidence identified in Step 2 as \"{**artifact_type**}\" and explain why it is more consistent with AI generation than with natural image formation. If the visual first impression in <first> suggested real, explain what it overlooked and why closer evidence supports fake. If it suggested fake, explain why closer evidence confirms it. Never mention that the target label, artifact region, or artifact type was provided in advance.*

Final Decision
Write the final verdict inside <answer> as exactly \"fake\".

STRICT RULES
*The visual first impression is \"{**first_prediction**}\", but the final answer should be \"fake\"; use the first impression only as context, do not justify it when it conflicts with fake, and write <think> so that the evidence supports <answer>fake</answer>.*
Step 1 must be exactly one short sentence. Steps 2 and 3 must each contain 2 to 3 complete sentences.
Steps 2 and 3 must use the same single visual evidence

Figure 10: **Artifact-guided fake response-generation prompt.** The red text marks the initial visual judgment, fake ground-truth label, artifact region, and artifact type provided during response generation. This prompt is used to construct RBI responses for manipulated images, grounding the target MLLM's reflection in the image-specific forensic evidence extracted by Gemini-2.5-Flash.



First GT

Case: *Real → Real*

<first>real</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression suggested "real," but I must rely on closer image-grounded evidence to confirm this. The image shows a natural and consistent lighting and shadow pattern that is typical of real photographs, with soft highlights and subtle variations in skin texture.

Step 2: Forensic evidence search:
*Upon closer inspection, **the lighting and shadows on the face and bottle are consistent with natural light sources, creating a realistic appearance.** The skin texture appears smooth with natural variations, which is common in real photographs. The bottle's reflection and lighting also match the lighting on the face, indicating a single light source.*

Step 3: Evidence interpretation:
The natural lighting and texture are plausible artifacts from real photographic phenomena, such as natural sensor noise and camera processing. Treating these as fake artifacts would be an overclaim, as they can arise from ordinary photographic conditions. The subtle variations in skin texture and lighting are typical of real images, making the impression of "real" valid. The consistent lighting on both the face and bottle suggests a real camera-captured image rather than AI generation, as AI often struggles to replicate such nuanced lighting and texture details accurately.
</think>
<answer>real</answer>



First GT

Case: *Real → Real*

<first>real</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression suggested "real," but I must rely on closer image-grounded evidence to confirm this. The image shows a natural, soft texture in the hair and skin, which is typical of real photographs. The subtle variations in lighting and shadowing on the face and hair are consistent with natural light and camera processing.

Step 2: Forensic evidence search:
*Upon closer inspection, the hair and skin exhibit a realistic, slightly uneven texture that is common in real photographs. **The soft, natural lighting and subtle shadows around the eyes and nose suggest a real camera-captured image**, as AI-generated images often have overly smooth textures and uniform lighting.*

Step 3: Evidence interpretation:
***The slight imperfections in the hair and skin**, such as the natural variation in texture and the subtle shadows, are plausible artifacts from physical lighting, optics, or natural sensor noise. Treating these as fake artifacts would be an overclaim, as they can arise from ordinary photographic phenomena like natural lighting and camera processing. The initial impression of "real" was correct because these details are consistent with real photography, not just an artifact of the impression itself.*
</think>
<answer>real</answer>



First GT

Case: *Fake → Real*

<first>fake</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression suggested "fake," but the final decision must be based on closer image-grounded evidence rather than that first impression alone.
Step 2: Forensic evidence search:
*Upon closer inspection, the image shows a consistent and natural skin texture with subtle variations in lighting and shadow that are typical of real photographs. **The lighting appears soft and diffused, which is often seen in natural lighting conditions, and the shadows on the face are soft and blended, indicating a real camera-captured image.***
Step 3: Evidence interpretation:
***Treating the observed skin texture and lighting as a fake artifact would be an overclaim** because they can plausibly arise from natural sensor noise and camera processing. The soft shadows and natural skin texture are common in real photographs, especially when taken with a high-quality camera. The slight imperfections and natural variations in lighting are not uncommon and can arise from ordinary photographic phenomena such as natural light and camera optics. The initial impression of "fake" may have overemphasized superficial cues, but the evidence of soft lighting and natural texture supports a real camera-captured interpretation.*
</think>
<answer>real</answer>



First GT

Case: *Fake → Real*

<first>fake</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression suggested "fake," but the final decision must be based on closer image-grounded evidence rather than that first impression alone.
Step 2: Forensic evidence search:
*Upon closer inspection, **the image shows a consistent and natural lighting pattern that is typical of real camera-captured photos**, with soft shadows and subtle highlights that suggest a natural light source, such as sunlight. The skin texture appears smooth and slightly grainy, which is consistent with natural sensor noise and camera processing.*
Step 3: Evidence interpretation:
*Treating the lighting and texture as fake artifacts would be an overclaim, as they can plausibly arise from natural lighting conditions and camera optics. The slight graininess and soft shadows are common in real photographs taken in bright, outdoor settings, and the smooth texture of the skin is a natural result of infant skin's delicate nature. **The impression of "fake" may have overemphasized superficial cues, such as the uniformity of the skin tone**, which can naturally occur in real images due to the child's age and the lighting conditions.*
</think>
<answer>real</answer>

Figure 11: **Qualitative inference examples on real images.** The examples include correct keeps (real → real) and corrective flips (fake → real). Ground-truth labels are shown only for analysis and are not provided during inference. SeeTR re-examines the initial visual judgment using natural lighting, shadows, texture variations, and camera characteristics, recognizing them as plausible photographic phenomena rather than overinterpreting them as manipulation artifacts. Highlighted phrases indicate the evidence supporting the final decision.



First GT

Case: *Fake* → *Fake*

<first>fake</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression is "fake," but I must rely on closer image-grounded evidence to confirm this.
Step 2: Forensic evidence search:
Upon closer inspection, the left earring shows a distinct distortion in the background, with a noticeable blur and unnatural blending around the edge, indicating AI generation.
Step 3: Evidence interpretation:
The background distortion around the earring is more consistent with AI generation than natural image formation because it lacks the smooth, organic blending typical of real images. The artificial earring appears to be superimposed, with a lack of depth and texture that is characteristic of AI-generated artifacts.
</think>
<answer>fake</answer>



First GT

Case: *Fake* → *Fake*

<first>fake</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression is "fake," but I must rely on closer image-grounded evidence to confirm this.
Step 2: Forensic evidence search:
Upon closer inspection, the left hand covering the forehead shows a distinct pattern of unnatural blending and pixelation, particularly around the fingers and the area where the hand meets the hair. The edges are overly smooth and lack the natural texture and shadowing of a real hand.
Step 3: Evidence interpretation:
The pixelation and blending indicate AI generation because the hand appears to *be artificially superimposed, with a lack of realistic skin texture and inconsistent lighting. The smoothness and uniformity in the hand's appearance suggest it was generated rather than naturally occurring, as seen in real images where skin texture and shadows are more varied and complex.*
</think>
<answer>fake</answer>



First GT

Case: *Real* → *Fake*

<first>real</first>
Step 1: Initial judgment awareness:
My initial visual impression suggested "real," but I must rely on closer image-grounded evidence to make a final decision.

Step 2: Forensic evidence search:
Upon closer inspection, the image shows a subtle mismatch in tone and color between the face and the neck, particularly around the jawline and the transition to the collar. The skin texture and lighting on the face appear smoother and more uniform compared to the neck, indicating a possible AI-generated artifact.

Step 3: Evidence interpretation:
The tone and color mismatch around the jawline and neck are more consistent with AI generation because natural images typically exhibit more variation in skin texture and lighting. AI often struggles to replicate the complex interplay of light and shadow in these areas, resulting in a more uniform and less realistic appearance. The smoother texture and lack of natural gradation in the neck area suggest the image was artificially created.
<answer>fake</answer>



First GT

Case: *Real* → *Fake*

<first>real</first>
<think>
Step 1: Initial judgment awareness:
My initial visual impression suggested "real," but I must rely on closer image-grounded evidence to make a final decision.

Step 2: Forensic evidence search:
Upon closer inspection, the beard and hair exhibit a painterly texture, particularly noticeable in the soft blending of edges and the uniformity of the hair strands, which lacks the natural variation and irregularity typically found in real human hair.

Step 3: Evidence interpretation:
The painterly texture in the beard and hair indicates AI generation because it shows a lack of natural variation and irregularity, which is more consistent with AI synthesis than with natural image formation. The uniformity and smoothness in the hair strands suggest a digital creation rather than a organic pattern.
</think>
<answer>fake</answer>

Figure 12: **Qualitative inference examples on fake images.** The examples include correct keeps (fake → fake) and corrective flips (real → fake). Ground-truth labels are shown only for analysis and are not provided during inference. SeeTR identifies localized artifacts such as background distortion, painterly hair texture, and face–neck appearance inconsistencies, using these cues to keep or flip the initial visual judgment toward the correct fake decision. Highlighted phrases indicate the evidence supporting the final decision.