REGION-TO-FAKE: PERCEPTUAL REINFORCEMENT LEARNING FOR FINE-GRAINED FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

While Multimodal Large Language Models (MLLMs) significantly enhance forgery perceptual image understanding, they still exhibit limited perceptual reasoning capabilities when applied to the specialized domain of fine-grained image forgery detection. To address these issues, we propose the Region-to-Fake (R2F) framework. Our approach leverages large-scale MLLMs combined with image-cropping tool invocation to distill multi-round image crops and bounding box coordinates directly into the Chain-of-Thought (CoT). This enables the student model to acquire internalized bounding box localization capabilities, effectively eliminating the reliance on external tools during inference. To further optimize the model's perceptual understanding of forgeries while mitigating the prolonged inference latency, we introduce Perceptual-Aware Reinforcement Learning with Verifiable Rewards (PA-RLVR). By leveraging the intrinsic tag-generation capabilities acquired during the cold-start phase, this strategy enables direct importance sampling at the tag-sequence level, which significantly enhances the performance of forgery localization. Extensive experiments demonstrate that our method achieves high performance and efficient inference latency across multiple Fake Image Detection and Localization benchmarks.

1 INTRODUCTION

Current paradigms for Fake Image Detection and Localization (FIDL) primarily rely on small vision models (Yan et al., 2025b;a; Wang et al., 2020) and Multimodal Large Language Models (MLLMs) (Wen et al., 2025; Tan et al., 2026; Jiang et al., 2026a). While these methods have driven significant progress, most focus on enhancing visual preprocessing pipelines or scaling up training data to improve Chain-of-Thought (CoT) reasoning. Consequently, they often fail to strengthen the intrinsic, fine-grained perceptual capabilities of the underlying models.To improve model perception of local image regions, recent research has explored the "thinking with images" paradigm (Zheng et al., 2025b). While this approach enhances an MLLM's fine-grained visual perception through external tool invocation, it introduces substantial inference latency. More importantly, as illustrated in Figure 1, it is tailored for general-purpose semantic objects. To bypass external tools, recent studies have successfully distilled the tool-calling CoT of Large Language Models (LLMs) directly into student models (Hu et al., 2024). However, these distillation efforts remain confined to the general domain.Unlike general objects with distinct semantic boundaries, forged regions in the forensic domain are intentionally designed to be imperceptible. Forgery clues manifest as subtle, low-level anomalies—such as region-level inconsistencies, blending artifacts, or unnatural noise distributions—rather than distinct semantic entities. We formulate this discrepancy, where models fail to precisely ground local artifacts and specific targets, as the **local referring gap**. For instance, if a single strand of hair is manipulated, current MLLMs typically localize the broader facial region rather than the exact artifact. Because existing general-purpose methods lack specialized fine-grained perception, they have not yet been successfully extended to FIDL (Du et al., 2025).

Motivated by these limitations, we propose the R2F framework, which categorizes training samples into fully AI-generated images and locally manipulated images for decoupled distillation. Our core innovation begins with the construction of a specialized dataset for cold-start: (1) for fully generated images, we combine LLMs and human annotation to generate descriptive datasets with multiple bounding boxes per image; (2) for locally manipulated images, we utilize the natively provided bounding boxes, augmenting them with LLM annotations and rigorous human review.

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.

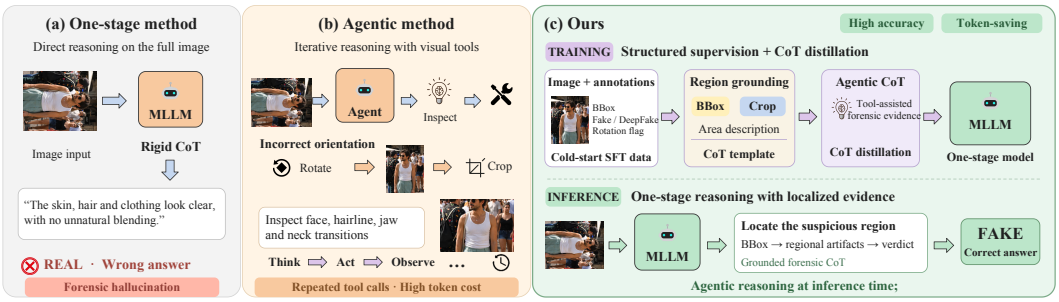

Figure 1: Comparison of fake image detection paradigms. One-stage methods are token-efficient but prone to rigid reasoning and forensic hallucination, while agentic methods improve accuracy through iterative visual operations at the cost of heavy token consumption.

Building upon this cold-start dataset, R2F distills crop-level bounding box coordinates directly into the model's CoT reasoning process during cold-start SFT. Following the paradigm of DeepSeek-R1 (Shao et al., 2024), we subsequently introduce Perceptual-Aware Reinforcement Learning with Verifiable Rewards (PA-RLVR). Unlike GRPO's token-level sampling (Shao et al., 2024) or GSPO's sequence-level sampling (Zheng et al., 2025a), we propose Tag-level Importance Sampling (IS-TS) tailored for structured multimodal reasoning. By leveraging the model's alternating outputs of `<think>...</think>` and `<bbox_2d>...</bbox_2d>` tags following the cold start, IS-TS estimates the importance of sentences enclosed within these specific tags. This strategy explicitly reinforces the MLLM's CoT outputs without requiring the strict token-level alignment typical of SFT. Consequently, this approach facilitates the precise extraction of intermediate localization trajectories and enhances overall forgery localization performance. Our main contributions are summarized as follows:

- We propose the R2Fframework to explicitly bridge the local referring gap in the forensic domain. By constructing a specialized cold-start dataset that decouples fully AI-generated and locally manipulated images, we successfully distill crop-level bounding box coordinates directly into the MLLM's intrinsic CoT reasoning process, thereby enhancing local anomaly perception without the latency of external tool invocation.
- We introduce PA-RLVR to further optimize the multimodal reasoning process. In contrast to standard token-level or sequence-level sampling, we propose IS-TS. By exploiting the model's alternating generation of `<think>` and `<bbox_2d>` tags, IS-TS dynamically estimates the importance of sentences within these specific tags. This bypasses the need for strict token-level SFT alignment and refines intermediate localization trajectories.
- We conduct comprehensive experiments across multiple benchmarks, demonstrating that R2Fcombined with PA-RLVR significantly outperforms existing methods. The results validate the effectiveness of our framework in advancing state-of-the-art FIDL performance.

2 Related Work

2.1 Fake Image Detection and Localization

Traditionally cast as an image classification task, early fake image detection methodologies predominantly relied on Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) to extract and distinguish forgery features (Team, 2026; Su et al., 2025; Zhu et al., 2025). For instance, Effort (Yan et al., 2025b) explored deep feature representations for identifying manipulated content. Furthermore, works like Imdl-benco (Ma et al., 2025) formally introduced the task of tampered image detection, while Forensichub (Du et al., 2025) established the first comprehensive benchmark specifically targeting Fake Image Detection and Localization (FIDL).

Recently, the research paradigm has shifted towards the interpretability of manipulated images, integrating Vision-Language Models (VLMs) to enhance detection. AntiFakePrompt (Chang et al., 2023) successfully introduced VLMs to the FID domain. However, Agent-based MLLMs often in-

cur substantial inference latency (Jiang et al., 2025; Zhao et al., 2026). For example, the Ivy-Fake framework (Jiang et al., 2026a) employs a Chain-of-Thought (CoT) paradigm that generates explanations prior to the final prediction, necessitating the output of a massive number of tokens. To alleviate this latency bottleneck, FakeVLM (Wen et al., 2025) proposed a "classify-then-explain" pipeline. While this resolves the latency issue, its natural language explanations remain overly vague and lack specificity, with extensive experiments revealing its limitations in detecting certain tampered images. Addressing the CoT overhead, Fake-HR1 (Jiang et al., 2026b) proposed a hybrid reasoning chain to specifically reduce the computational burden of CoT in FID scenarios.

2.2 Forgery Perception and Explainability

Considerable research has been dedicated to establishing explainable forgery perception. In this vein, Ivy-Fake (Jiang et al., 2026a) introduced the first jointly trained dataset for image and video forgery. Subsequent frameworks, such as FakeBench (Li et al., 2025a), AIGI-Holmes (Zhou et al., 2025), and Veritas (Tan et al., 2026), proposed explainable detection pipelines tailored for various categories and multiple granularities of AI-generated images. Additionally, Forensic-Chat (Lin et al., 2025) adopted a "reason-then-detect" schema to guide model perception, while AgentFox (Yu et al., 2026) introduced an agent-based approach for generated content detection.

Despite these advancements, existing approaches suffer from insufficient fine-grained explainability. They fail to establish a robust correlation between the reasoning process and the final classification, and often struggle to accurately localize the forged regions. Although FakeVLM-R1 (Zhu et al., 2026b) and SIDA (Huang et al., 2025) explores the use of bounding boxes in the AIGC detection domain, it still suffers from the local referring gap. Furthermore, its proposed dataset fails to address the localization of local manipulations and document text forgeries within the FIDL task. Additionally, while the "thinking with images" paradigm (Zheng et al., 2025b) has been widely adopted to enhance visual grounding, its high inference latency and practical viability (Li et al., 2026) remain questionable. Recently, ZwZ (Wei et al., 2026) demonstrated that MLLM forgery perception could be improved without explicit tool invocation. However, its reliance on direct Question-Answering (QA) without a coherent reasoning chain struggles to provide compelling evidence to verify its conclusions.

To overcome these limitations, we propose R2F that distills multi-round image-cropping tool interactions into a single-round CoT reasoning process. by leveraging reinforcement learning (RL), our method strengthens the fine-grained perceptual understanding and localization capabilities of MLLMs, while mitigating the input-token overhead caused by cropped image inputs. As a result, it better aligns interpretable reasoning with efficient and accurate forgery detection.

3 Method

3.1 Region-to-Fake

It is well-acknowledged that MLLMs typically take an image I and a question Q to construct a specific prompt, generating a sequence of tokens T autoregressively. For Forgery Image Detection tasks, performance is ultimately evaluated based on specific output tokens. However, the length of T is the primary bottleneck for inference latency. Previous "Think with Images" paradigms rely on multi-turn dialogue contexts, denoted as $T = \{t_1, t_2, ..., t_n\}$, which incur substantial inference latency. Furthermore, conventional pure "think-answer" schemes suffer from inadequate fine-grained understanding of local regions. Their text-only CoT reasoning is frequently vague, lacking visually grounded interpretability to convincingly support forgery detection.

To tackle these limitations, we propose the Region-to-Fake (R2F) cold-start data distillation scheme. R2F distills the fine-grained local region comprehension derived from a complex multi-model pipeline into a single, efficient MLLM. By explicitly integrating bounding box coordinates into the reasoning process and compressing multi-step tool-call outputs into a single-turn think-answer format via Supervised Fine-Tuning (SFT), we endow the model's CoT with robust Fine-Grained Forgery Detection capabilities, which are further stimulated via subsequent Reinforcement Learning (RL).

Tag-level Weighted Loss. In our framework, the model generates an interleaved sequence of textual reasoning and spatial grounding, ultimately concluding with a task-specific answer. The output sequence T is composed of three major tagged spans: reasoning tokens T_{think} enclosed by `<think>`, spatial grounding tokens T_{bbox} enclosed by `<bbox_2d>`, and final prediction tokens T_{answer} enclosed by `<answer>`. Standard autoregressive training assigns an identical loss weight to all tokens, which may dilute the supervision signals for critical spatial coordinates and the final task-specific prediction. This issue is particularly important in our setting, where the correctness of the generated bounding box directly affects visual grounding quality, while the final answer determines the evaluation metric. To enforce strict accuracy on both spatial grounding and final prediction, we introduce a *Tag-level Weighted Loss*, denoted as **Tag-Loss**. Specifically, we assign different importance weights to tokens according to their enclosing tag segment. Given an input image I, question Q, and target sequence $T = \{t_i\}_{i=1}^{|T|}$, the training objective is defined as:

$$\mathcal{L}_{\text{Tag-Loss}} = -\sum_{i=1}^{|T|} w_i \log P_\theta(t_i \mid I, Q, t_{<i}), \tag{1}$$

where w_i denotes the dynamic weight applied to the i-th token t_i. The token weight is determined by its tag role:

$$w_i = \begin{cases} w_{\text{think}}, & t_i \in T_{\text{think}}, \\ w_{\text{bbox}}, & t_i \in T_{\text{bbox}}, \\ w_{\text{answer}}, & t_i \in T_{\text{answer}}, \\ 1, & \text{otherwise}, \end{cases} \tag{2}$$

where w_{think} is set to 1 by default, while w_{bbox} and w_{answer} are used to emphasize the bounding box coordinates and the final task-specific answer, respectively. Tokens outside the explicitly defined tag segments are kept with the default language modeling weight of 1. In practice, we set $w_{\text{bbox}} > w_{\text{think}}$ and $w_{\text{answer}} > w_{\text{think}}$, encouraging the model to preserve general reasoning ability while placing stronger optimization pressure on spatial localization and final prediction accuracy. The ablation study of weight can be seen in Appendix.

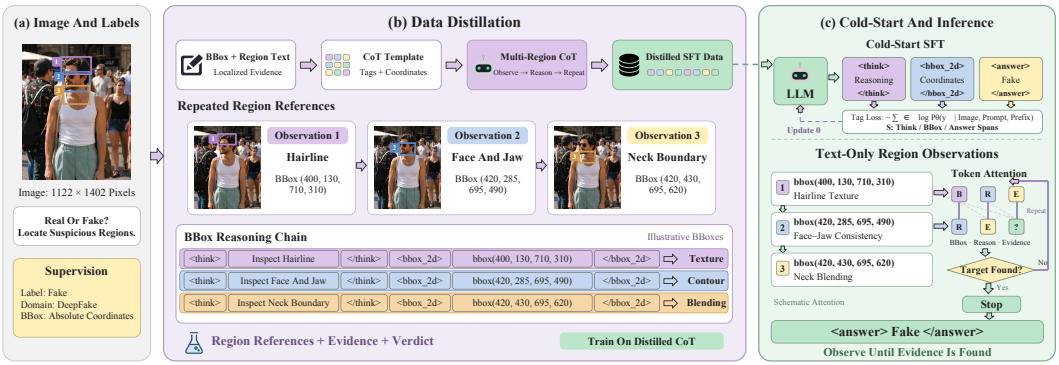

Figure 2: Example of cold-start data construction. We leverage local bbox annotations when available; otherwise, the model explores regions autonomously.

Cold Start SFT Training Data. To initialize the policy model with robust foundational capabilities across diverse domains, we construct a high-quality cold-start multi-domain dataset based on the training splits of OpenMMSec Du et al. (2026) and Ivy-Fake Jiang et al. (2026a). Specifically, as shown in Figure 2, we integrate data from multiple specialized sources and further enrich OpenMMSec with additional samples from M3DSynth Zingarini et al. (2024) and M6Doc Cheng et al. (2023), which supplement the medical and document forensics domains, respectively. This curation process yields 40,935 high-quality instruction-tuning samples for the cold-start SFT phase and 66,324 multi-turn trajectory samples for the subsequent PA-RLVR stage. Such progressive data scheduling enables the model to first acquire basic multimodal forensic capabilities before further exploring complex reasoning trajectories for robust deepfake detection and security-oriented localization. For samples without bounding-box annotations, we constrain the generated CoT to contain only one or two `bbox` tags. If the initial CoT contains more than two `bbox` tags, we resample

it; for cases that consistently exceed this limit, we retain only two `bbox` tags and append the final answer after selecting the CoT whose prediction matches the ground-truth label. For IMDL-domain samples with bounding-box annotations, we distill the final CoT following the R2F scheme. The detailed distillation procedure and prompts are provided in Appendix.

3.2 PA-RLVR

Tag-Sequence Importance Sampling. Existing reinforcement learning methods, such as GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025a), usually compute importance sampling ratios at either the token level or the whole-sequence level. However, in our structured reasoning format, different tags often have highly imbalanced lengths. For example, <bbos_2d> segments contain dense visual evidence, while <think> segments are usually much shorter and focus on concise reasoning. Directly applying sequence-level importance sampling may therefore bias the policy update toward longer segments and weaken the contribution of shorter but critical reasoning tags.

To mitigate this issue, we propose **Tag-Sequence Importance Sampling (TS-IS)**, which computes the importance ratio independently within each tag segment. Given a trajectory τ_i composed of S_i valid tag segments, we denote the j-th segment as $\tau_i^{(j)}$ with length $|\tau_i^{(j)}|$. Let $r_{i,t}$ denote the token-level log-ratio:

$$r_{i,t} = \log \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}. \tag{3}$$

The normalized importance ratio for the j-th tag segment is:

$$w_i^{(j)} = \exp\left(\frac{1}{|\tau_i^{(j)}|} \sum_{t \in \tau_i^{(j)}} r_{i,t}\right). \tag{4}$$

We then average the ratios over all tag segments to obtain the final trajectory-level importance weight:

$$w_i^{\text{TS-IS}} = \frac{1}{S_i} \sum_{j=1}^{S_i} w_i^{(j)}. \tag{5}$$

This tag-wise normalization reduces the dominance of long observation segments and yields a more balanced policy update across perception, reasoning, and answer generation.

Reward Design. To supervise the structured generation process, we decompose the reward into three tag-specific components corresponding to reasoning, localization, and answer prediction. For a generated trajectory, the final reward is defined as:

$$R = \lambda_1 R_{\text{think}} + \lambda_2 R_{\text{bbox_2d}} + \lambda_3 R_{\text{answer}}, \tag{6}$$

where R_{think}, R_{bbox}, and R_{answer} denote the rewards for the <think>, <bbox_2d>, and <answer> segments, respectively. The coefficients λ_1, λ_2, and λ_3 balance the contribution of each component.

The reasoning reward R_{think} encourages the model to generate valid and coherent intermediate reasoning. The localization reward R_{bbox} evaluates all bounding boxes generated inside the <bbox_2d> tags. We use the IoU score as a soft localization reward, denoted as SoftIoU, since it provides a continuous-valued signal in $[0, 1]$ instead of a binary matching result. Given a predicted box $\hat{b}$ and a ground-truth box b, it is computed as:

$$\text{SoftIoU}(\hat{b}, b) = \frac{\text{Area}(\hat{b} \cap b)}{\text{Area}(\hat{b} \cup b) + \epsilon}, \tag{7}$$

where ϵ is a small constant for numerical stability.

Given M valid predicted boxes $\{\hat{b}_m\}_{m=1}^{M}$ and N ground-truth boxes $\{b_n\}_{n=1}^{N}$, each predicted box is matched to the ground-truth box with the maximum SoftIoU. The localization reward is then defined as:

$$R_{\text{bbox_2d}} = \frac{1}{M} \sum_{m=1}^{M} \max_{1 \le n \le N} \text{SoftIoU}(\hat{b}_m, b_n). \tag{8}$$

The answer reward R_{answer} measures the correctness of the final response against the ground-truth answer. By combining these rewards, our objective jointly optimizes reasoning, spatial grounding, and final task performance.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

In this section, we conduct extensive experiments and comprehensive ablation studies to evaluate the effectiveness of our proposed framework. The detailed descriptions of the evaluation datasets and the specific hyperparameter configurations utilized during the training phases are deferred to Appendix. Unless otherwise specified, R2F refers to the R2F-8B model in the following experiments.

4.2 MAIN RESULT

Table 1: Performance comparison (Acc.) on benchmarks. All results are averaged.

Method	Size	In-domain						Out-of-domain		
		GenImage	AIGCD.	CommunityForensics	AIGIBench	DiffusionForensics	MMSec$_{ID}$	LOKI	Chameleon	MMSec$_{OOD}$
General-domain MLLMs										
Qwen2.5-VL	3B	70.96	62.06	49.10	55.56	68.84	65.70	60.62	51.77	<u>69.50</u>
	7B	52.29	53.05	85.66	59.21	41.27	48.40	48.80	61.16	41.70
Qwen3-VL	4B	70.55	68.70	89.71	74.21	55.14	54.46	79.41	<u>74.18</u>	53.25
	8B	55.45	58.72	87.08	65.92	45.20	50.37	59.83	71.67	45.74
InternVL3.5	4B	48.62	52.34	25.79	46.85	15.55	13.58	50.43	13.30	31.74
	8B	55.18	58.47	59.88	52.36	1.73	8.34	56.79	3.63	34.92
FID-domain MLLMs										
FakeShield (ICLR'25)	22B	88.39	73.50	69.24	69.22	84.86	84.86	62.61	54.24	68.58
FakeVLM (NeurIPS'25)	7B	**99.43**	<u>91.53</u>	92.09	62.31	71.51	62.10	<u>84.30</u>	62.85	67.79
Ivy-xDetector	3B	96.32	88.86	<u>94.60</u>	<u>77.49</u>	**93.47**	58.00	71.18	73.17	60.58
Veritas (ICLR'26)	8B	98.04	88.67	89.12	76.09	66.46	61.34	68.88	59.57	68.30
R2F	8B	<u>99.20</u>	**93.25**	**96.73**	**82.36**	<u>92.80</u>	**89.22**	**90.14**	**79.02**	**82.59**

Comparison between general-domain MLLMs and FID-domain MLLMs. As shown in Table 1, general-domain MLLMs exhibit limited and unstable performance on fake image detection benchmarks, especially on forensic-oriented datasets. Although models such as Qwen3-VL-4B achieve reasonable results on several subsets, their performance varies significantly across benchmarks and does not consistently generalize to out-of-domain scenarios. For example, Qwen2.5-VL-7B obtains strong performance on CommunityForensics but performs poorly on GenImage, DiffusionForensics, and MMSec$_{OOD}$. In contrast, FID-domain MLLMs achieve substantially stronger and more consistent performance, demonstrating the importance of domain-specific training for fake image detection and localization.

Table 2: Comparison with baseline methods on OpenMMsec across four domains.

Method	Venue	DeepFake	AIGC	IMDL	Doc	Avg
Resnet (He et al., 2016)	CVPR'16	67.8	74.2	77.3	71.7	72.7
EfficientNet (Tan et al., 2019)	ICML'19	43.3	63.0	51.4	61.9	54.9
CapsuleNet (Sabour et al., 2017)	NeurIPS'17	62.0	75.4	77.2	72.3	71.7
SegFormer (Xie et al., 2021)	NeurIPS'21	80.7	**85.9**	81.7	<u>73.4</u>	<u>80.4</u>
Swin (Liu et al., 2021)	ICCV'21	79.0	<u>85.4</u>	82.9	72.0	79.8
Trufor (Guillaro et al., 2023)	CVPR'23	72.2	82.5	80.5	72.3	76.9
UnivFD (Ojha et al., 2023)	CVPR'23	54.2	79.8	70.9	62.7	66.9
Effort (Yan et al., 2025b)	ICML'25	**85.0**	81.9	<u>83.7</u>	69.6	80.1
Mesorch (Zhu et al., 2025)	AAAI'25	75.7	81.4	79.9	72.7	77.4
CO-SPY (Cheng et al., 2025)	CVPR'25	72.2	83.3	76.0	70.0	75.4
R2F	-	<u>84.3</u>	84.0	<u>82.1</u>	**80.0**	**82.6**

R2F achieves consistently strong performance. Compared with existing FID-domain MLLMs, R2F achieves the best overall performance across both in-domain and out-of-domain benchmarks. In the in-domain setting, R2F obtains the highest accuracy on AIGCD., CommunityForensics, AIGIBench, and MMSec$_{ID}$, while remaining highly competitive on GenImage and DiffusionForensics. More importantly, R2F shows clear advantages under distribution shift, achieving the best

results on all three out-of-domain benchmarks, including 90.14 on LOKI, 79.02 on Chameleon, and 82.59 on MMSec$_{OOD}$.

Comparison across multi domains. We compare R2F with existing baseline methods on OpenMMsec. As shown in Table 2, R2F achieves the best average performance across all domains, reaching 82.6 Avg accuracy and outperforming the strongest baseline SegFormer by 2.2 points. Notably, R2F shows clear superiority on the Doc domain, improving over the second-best method by 6.6 points, which indicates stronger robustness for document-oriented forgery detection. Meanwhile, R2F also maintains competitive performance on DeepFake, AIGC, and IMDL, achieving either the best or second-best results across these domains. The more performance of localization can be seen in Appendix **??**.

Table 3: Robust Analysis of Image Quality Degradation on $MMSec_{OOD}$ (Du et al., 2026). Results of the compared models are reported from DeFakerOne (Team, 2026).

Method		FFDN	Mesorch	Effort	ForensicsAdapter	ForensicMOE	FakeShield	R2F
Gaussian Blur	0.5	42.84	51.78	59.30	57.65	55.70	62.86	**81.76**
	1.5	48.68	49.07	56.55	55.75	53.63	65.57	**80.95**
	2.5	50.62	47.30	54.03	53.77	52.95	65.74	**78.09**
JPEG Compression	75	39.95	51.78	58.68	56.67	51.35	62.94	**82.23**
	85	39.38	52.00	57.13	56.37	51.25	63.46	**80.96**
	95	40.01	52.25	59.53	57.15	53.33	62.74	**77.67**
Resize	128	49.48	50.53	57.18	54.02	52.50	52.13	**79.09**
	256	45.32	49.32	58.43	57.25	54.15	50.85	**78.84**
	512	40.51	50.98	59.55	58.17	53.68	51.35	**76.21**

In-the-wild generalization. Table 4 evaluates the generalization ability of different methods on two challenging real-world benchmarks, LOKI and Chameleon. Compared with existing AIGC detectors and general-purpose MLLMs, R2F consistently achieves stronger and more balanced performance across real and fake samples. On LOKI, R2F obtains 90.1% accuracy and 89.5 F1, outperforming prior methods under the in-the-wild setting. On Chameleon, where many baselines suffer from severe real/fake bias, R2F still maintains competitive real accuracy and fake accuracy, achieving 79.0% accuracy and 78.6 F1. Overall, R2F reaches the best average performance with 84.6% accuracy and 84.1 F1, demonstrating robust cross-domain generalization and reduced reliance on dataset-specific artifacts. We further compare R2F with additional state-of-the-art detectors on the LOKI benchmark in Appendix.

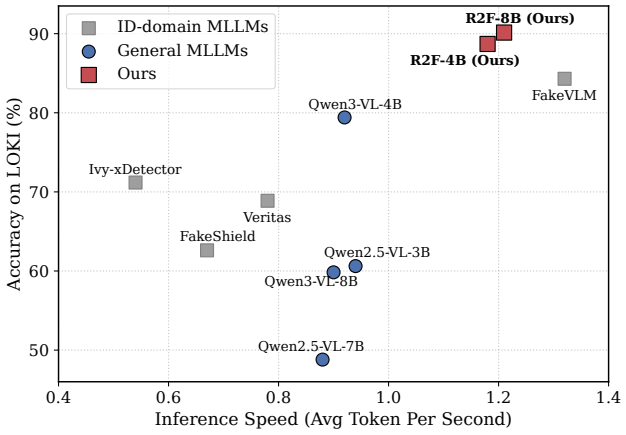

Figure 3: Efficiency performance trade-off on LOKI.

Efficiency performance trade-off. Figure 3 compares different MLLMs on LOKI in terms of inference speed and detection accuracy. General-domain MLLMs show limited general-domain

forensic capability, where Qwen2.5-VL and Qwen3-VL variants achieve relatively low or unstable accuracy despite comparable inference speeds. FID-domain MLLMs generally perform better, suggesting that domain-specific MLLM's training is important for FID task. Notably, FakeVLM achieves a competitive speed and is slightly faster than our model, but its accuracy remains clearly lower, indicating that higher inference efficiency alone is insufficient for robust detection.

4.3 Ablation Study

Table 4: In-the-Wild Generalization: Detection Evaluation on Real-World Chameleon (Yan et al., 2025a) and LOKI (Ye et al., 2025). Other metrics are source from EvoGuard (Zhu et al., 2026a) for LOKI evaluation.

Method	Venus	LOKI		Chameleon		Overall	
		R-Acc/F-Acc/Acc	F1	R-Acc/F-Acc/Acc	F1	R-Acc/F-Acc/Acc	F1
AIDE (Yan et al., 2025a)	ICLR'25	85.7/61.3/73.7	72.0	95.1/26.8/66.7	-	90.4/44.1/70.2	-
ForensicMOE (Fang et al., 2025)	ICCV'25	77.6/72.3/74.9	77.1	98.8/5.9/58.9	11.0	88.2/39.1/66.9	44.1
Qwen3-VL-2B (Bai et al., 2025a)	Arxiv'25	53.3/80.1/66.7	66.7	99.3/10.5/61.2	18.8	76.3/45.3/64.0	42.8
Qwen3-VL-4B (Bai et al., 2025a)	Arxiv'25	49.1/92.2/79.4	70.7	95.8/45.4/74.2	60.2	72.5/68.8/76.8	65.5
Qwen3-VL-8B (Bai et al., 2025a)	Arxiv'25	84.2/87.2/59.8	88.1	97.6/37.1/71.7	52.9	90.9/62.2/65.8	70.5
FakeVLM (Wen et al., 2025)	NeurIPS'25	74.3/88.9/84.3	86.1	35.2/99.7/62.8	69.7	54.8/94.3/73.6	77.9
FakeShield (Xu et al., 2025)	ICLR'25	47.5/75.6/61.6	71.5	73.4/28.7/54.2	35.0	60.5/52.2/57.9	53.3
Veritas (Tan et al., 2026)	ICLR'26	56.1/77.6/68.9	74.8	35.5/91.6/59.6	66.0	45.8/84.6/64.3	70.4
Ivy-xDetector (Jiang et al., 2026a)	ICMR'26	59.0/79.5/71.2	76.6	75.2/63.9/70.3	64.9	67.1/71.1/70.8	70.8
R2F	-	93.2/87.0/90.1	89.5	83.2/74.8/79.0	78.6	88.2/80.9/84.6	84.1

Robustness Analysis on Image Quality Degradation. To evaluate the generalization and robustness of our proposed method under real-world distortions, we conduct an ablation study by applying three common types of image quality degradation—Gaussian Blur, JPEG Compression, and Resize—with varying intensities on the $MMSec_{OOD}$ dataset. As shown in Table 3, our method (R2F) consistently outpeforms all compared domain-specific and multi-modal security models by a significant margin across all degradation levels. Notably, even under severe distortions, our method maintains a high accuracy of over 76%, while the performance of existing methods drops substantially (mostly hovering around 50%-65%).

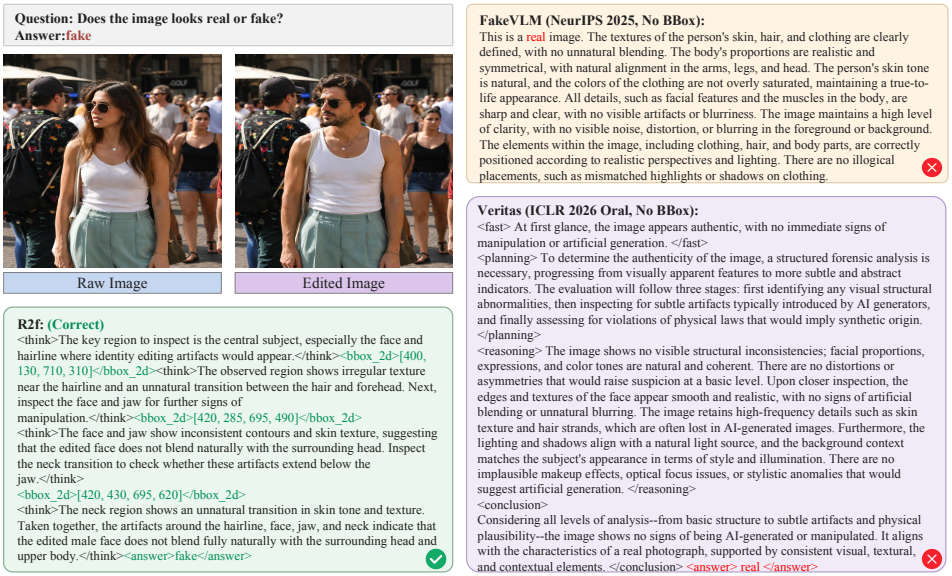

Figure 4: Reasoning comparisons between R2F and existing FID-domain MLLMs. Different from the horizontally rotated input in Figure 1, this case uses an upright image.

Table 5: Ablation Study of PA-RLVR.

Method	LOKI	Chameleon
Zero-shot (Qwen3-VL-8B)	59.83	71.67
SFT	62.15	72.84
+ Tag-Loss (Ours)	68.40	74.50
GRPO	83.55	76.20
+ Sequence Sampling	85.12	76.65
+ TS-IS (PA-RLVR, Ours)	**90.14**	**79.02**

Table 6: Ablations on the reward weights.

λ_1	λ_2	λ_3	LOKI	Chameleon
0.1	0.1	0.9	91.21	77.10
0.1	0.2	0.7	88.64	79.55
0.1	0.3	0.6	90.14	79.02

Table 7: Ablations on different base models and model sizes.

Model	LOKI	Chameleon
Qwen2.5-VL-7B	90.09	77.50
InternVL3.5-8B	89.60	80.12
Qwen3-VL-4B	88.87	78.65
Qwen3-VL-8B	90.14	79.02

Component-wise Ablation Study. Experimental results in Table 5 demonstrate the progressive efficacy of our framework, where standard SFT yields only marginal gains over the Zero-shot base, whereas incorporating our proposed *Tag-Loss* substantially boosts performance, and further integrating our core process-aware *TS-IS (PA-RLVR)* token-and-sequence optimization triggers a decisive performance leap over vanilla GRPO and Sequence Sampling to achieve the best results on both benchmarks. The necessity analysis of TS-IS is provided in Appendix.

Effects of Reward Weights. As shown in Table 6, to investigate how different components of the reward function affect the model's performance, we conduct an ablation study on the reward weights (λ_1, λ_2, λ_3) on the LOKI and Chameleon datasets. As illustrated in Table 6, the model achieving the optimal trade-off across benchmarks is configured with $\lambda_1 = 0.1, \lambda_2 = 0.3, \lambda_3 = 0.6$ (Default). Specifically, we observe that assigning a higher weight to the final task-specific correctness ($\lambda_3 = 0.9$) significantly boosts the performance on LOKI (91.21%) but leads to a noticeable drop on Chameleon (77.10%), suggesting a potential over-optimization on a single objective.

Impact of Base Models and Sizes. As summarized in Table 7, our framework demonstrates strong compatibility across various popular backbones, including the Qwen Bai et al. (2025a) and InternVL series Wang et al. (2025). Comparing `Qwen3-VL-4B` with `Qwen3-VL-8B` (Default) confirms that scaling up the parameter size consistently yields performance gains across both datasets ($+1.27\%$ on LOKI and $+0.37\%$ on Chameleon), which aligns with standard scaling laws. Therefore, we select `Qwen3-VL-8B` as our default backbone to ensure robust cross-domain generalization.

More Ablation Studies. We provide additional ablation studies in the appendix, including the hyperparameter ablation of tag-level loss weights, the effect of tag-level loss on RL training stability, and the ablation study on training data.

4.4 Qualitative Analysis

As shown in Figure 4, existing methods such as FakeVLM and Veritas mainly rely on global visual descriptions and incorrectly predict the edited image as real. In contrast, R2F explicitly grounds its analysis on the manipulated facial region, including the face, hairline, jaw, and neck transition.

5 Conclusion

In this paper, we propose R2F, a novel data distillation framework for addressing fine-grained forgery detection with MLLMs. By directly generating bounding boxes within a single-round CoT process, R2F improves the practical usability of MLLMs for forensic detection. To further strengthen this process, we introduce PA-RLVR, which stabilizes model training through tag-level sentence-importance sampling. Extensive experiments demonstrate the superior performance of our method on FID tasks with low inference latency. Meanwhile, the model also exhibits promising localization ability, showing its potential for practical fine-grained forgery analysis.

Due to computational resource constraints, we currently validate the effectiveness of PA-RLVR mainly in the FID domain. Although our experiments show that the proposed model also achieves strong localization performance, localization is not the primary focus of this work. In addition, our experiments are limited to base models with fewer than 8B parameters. We plan to explore larger-scale MLLMs in future work to investigate whether stronger model capacity can further improvelocalization, and detection performance.

AI USE STATEMENT

In this work, we used generative AI tools to generate and refine chain-of-thought training data and region-level descriptions, and to assess prediction accuracy and reasoning–prediction consistency when constructing preference pairs for DPO training. We have not used generative AI tools for other tasks requiring disclosure, and the remaining required-disclosure categories are not applicable to this work. Additionally, we used generative AI tools for translation, grammar correction, and small rhetorical edits intended to preserve scientific meaning. We have reviewed all AI-assisted work. Our responsibility includes verifying the accuracy of generated annotations and ensuring that language edits preserve the original claims, numerical results, uncertainty, and limitations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies image forgery detection and localization to support the verification of digital content. It does not involve human-subject studies or interventions. We use existing datasets for research purposes and recognize the importance of respecting their licenses, access conditions, and privacy requirements, particularly for images containing faces or documents. Forensic predictions can produce false positives and false negatives, and generated explanations may be inaccurate even when a prediction is correct. Accordingly, model outputs should be treated as supporting evidence requiring independent verification, especially in consequential settings, rather than as definitive judgments of authenticity. Performance may also vary across image sources, manipulation types, and demographic groups; our benchmark results do not establish demographic fairness or universal reliability. Although this research aims to improve detection, insights into forensic cues could also inform attempts to evade detectors. These limitations and potential misuse should be considered when deploying or releasing related models and data.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility and support the research community, we provide detailed descriptions of the experimental setup and hyperparameter settings in the appendix. These details are intended to help researchers reproduce our results and build upon our work.

REFERENCES

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

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

You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15138–15147, June 2023.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13455–13465, 2025.

Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pp. 422–426, Beijing, China, Jul 2013. IEEE. ISBN 978-1-4799-1043-4. doi: 10.1109/ChinaSIP.2013.6625374. URL http://ieeexplore.ieee.org/document/6625374/.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Ji-Zhe Zhou. Forensichub: A unified benchmark & codebase for all-domain fake image detection and localization. *arXiv preprint arXiv:2505.11003*, 2025.

Bo Du, Xiaochen Ma, Xuekang Zhu, Zhe Yang, Chaogun Niu, Chenfan Qu, Mingqi Fang, Zhenming Wang, Jingjing Liu, Jian Liu, and Ji-Zhe Zhou. Can we build a monolithic model for fake image detection? sica: Semantic-induced constrained adaptation for unified-yet-discriminative artifact feature space reconstruction, 2026. URL https://arxiv.org/abs/2602.06676.

Mingqi Fang, Ziguang Li, Lingyun Yu, Quanwei Yang, Hongtao Xie, and Yongdong Zhang. Forensic-moe: Exploring comprehensive synthetic image detection traces with mixture of experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17772–17782, October 2025.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20606–20615, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024.

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

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 893–903, 2023.

Changjiang Jiang, Fengchang Yu, Haihua Chen, Wei Lu, and Jin Zeng. Tabdsr: Decompose, sanitize, and reason for complex numerical reasoning in tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3172–3196, 2025.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, ICMR '26, pp. 2438–2447. Association for Computing Machinery, 2026a. ISBN 9798400726170. doi: 10.1145/3805622.3810615. URL https://doi.org/10.1145/3805622.3810615.

Changjiang Jiang, Xinkuan Sha, Fengchang Yu, Jingjing Liu, Jian Liu, Mingqi Fang, Chenfeng Zhang, and Wei Lu. Fake-hr1: Rethinking reasoning of vision language model for synthetic image detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10482–10486, 2026b. doi: 10.1109/ICASSP55912.2026.11462736.

Haobin Li, Yutong Yang, Yijie Lin, Dai Xiang, Mouxing Yang, and Xi Peng. Reliable thinking with images. *arXiv preprint arXiv:2602.12916*, 2026.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. Fakebench: Probing explainable fake image detection via large multimodal models. *IEEE Transactions on Information Forensics and Security*, 20:8730–8745, 2025a. doi: 10.1109/TIFS.2025.3597211.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *arXiv preprint arXiv:2505.12335*, 2025b.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Junyan Ye, Ke-Yue Zhang, Yue Zhou, Peng Jin, Bin Li, Taiping Yao, and Shouhong Ding. Seeing before reasoning: A unified framework for generalizable and explainable fake image detection. *arXiv preprint arXiv:2509.25502*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157:110828, 2025.

Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37:134591–134613, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24480–24489, 2023.

Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5937–5946, 2023.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 694–702, 2025.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7024–7032, 2025.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *ICLR*, 2026.

Mingxing Tan, Q Efficientnet Le, et al. Rethinking model scaling for convolutional neural networks. In *Proceedings of the International conference on machine learning, Long Beach, CA, USA*, volume 15, 2019.

GuangJian Team. Venus-defakerone: Unified fake image detection & localization. *arXiv preprint arXiv:2605.14091*, 2026.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pp. 8692–8701, 2020. doi: 10.1109/CVPR42600.2020.00872.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *ICCV*, pp. 22388–22398, 2023. doi: 10.1109/ICCV51070.2023.02051.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *NeurIPS*, 2025.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lv7PjbcaMi.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *ICLR*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *ICLR*, 2025a.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Forty-second International Conference on Machine Learning*, 2025b.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. In *ICLR*, 2025.

Yangxin Yu, Yue Zhou, Bin Li, Kaiqing Lin, Haodong Li, Jiangqun Ni, and Bo Cao. Agentfox: Llm agent-guided fusion with explainability for ai-generated image detection. *arXiv preprint arXiv:2603.23115*, 2026.

Ming Zhao, Wenhui Dong, Yang Zhang, XiangZheng, Zhonghao Zhang, Zian Zhou, YUNZHI GUAN, Liukun Xu, Wei Peng, Zhaoyang Gong, Zhicheng Zhang, Dachuan li, Xiaosheng Ma, Yuli Ma, Jianing Ni, Changjiang Jiang, Lixia Tian, Chen Qixin, Xia Kaishun, Pingping Liu, Tongshun Zhang, ZhiqiangLiu, Zhongan Bi, Chenyang Si, Tiansheng Sun, and Caifeng Shan. Spinebench: A clinically salient, level-aware benchmark powered by the spinemed-450k corpus. In *The Fourteenth International Conference on Learning Representations*, 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, 2025a. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025b.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

Chenyang Zhu, Maorong Wang, Jun Liu, Ching-Chun Chang, and Isao Echizen. Evoguard: An extensible agentic rl-based framework for practical and evolving ai-generated image detection. *arXiv preprint arXiv:2603.17343*, 2026a.

Leqi Zhu, Junyan Ye, Kaiqing Lin, Zhiyuan Yan, Conghui He, and Weijia Li. Fakevlm-r1: Internalizing physical laws via cot for synthetic image detection, 2026b. URL https://arxiv.org/abs/2605.30062.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image, 2023.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11022–11030, 2025.

Giada Zingarini, Davide Cozzolino, Riccardo Corvi, Giovanni Poggi, and Luisa Verdoliva. M3dsynth: A dataset of medical 3d images with ai-generated local manipulations, 2024. URL https://arxiv.org/abs/2309.07973.

A LICENSE

All datasets and models used in this work are available for research purposes and public available. We further declare that all models developed by us are used strictly for research purposes only.

B SOCIAL IMPACT

Our work aims to advance reliable AIGC and DeepFake detection for research and forensic analysis. However, we acknowledge that AIGC detectors may also be misused to train adversarial generation models that evade detection or improve the realism of DeepFake content. Such misuse could potentially weaken existing detection systems and amplify risks related to misinformation, identity fraud, and malicious content manipulation.

We emphasize that our study is conducted strictly for research purposes and follows commonly accepted ethical guidelines in the multimedia forensics and AI safety communities. The models and resources developed in this work are intended to support trustworthy media analysis, improve robustness against synthetic content, and facilitate responsible academic research.

C EXPERIMENT SETUP

Training Setup. The model training consists of two stages: Cold Start SFT and PA-RLVR. In the Cold Start SFT stage, we set the learning rate to 1×10^{-5}, batch size to 64, warmup ratio to 0.05,

gradient accumulation steps to 8, the ViT learning rate to 1×10^{-6}, and the maximum number of image tokens to 2048. In the PA-RLVR stage, we use the same learning rate as in the SFT stage, and set the number of GRPO sampled responses to $n = 8$.

Baselines. We evaluate three categories of models: Small Vision Models, General-domain MLLMs, and FID-domain MLLMs. The General-domain MLLMs include Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), and InternVL3.5 (Wang et al., 2025). The FID-domain MLLMs include Veritas (Tan et al., 2026), FakeShield (Xu et al., 2025), Ivy-xDetector (Jiang et al., 2026a), and FakeVLM (Wen et al., 2025).

Benchmarks. We evaluate our model on both in-domain and out-of-domain benchmarks. The in-domain benchmarks include GenImage (Zhu et al., 2023), AIGCDetectionBenchmark (AIGCD.) (Zhong et al., 2023), AIGIBench (Li et al., 2025b), DiffusionForensics (Wang et al., 2023), and the OpenMMSec in-domain test set (Du et al., 2026). The out-of-domain benchmarks include LOKI (Ye et al., 2025), Chameleon (Yan et al., 2025a), and the OpenMMSec out-of-domain test set (Du et al., 2026).

D Relationship between Tag-Loss and Dynamic Fine-Tuning

Our Tag-Loss is conceptually related to recent Dynamic Fine-Tuning (DFT) (Wu et al., 2026), which revisits SFT from a reinforcement learning perspective and improves generalization by dynamically rescaling token-level objectives with model-assigned probabilities. Instead of treating all target tokens uniformly as in standard SFT, both DFT and our method modify the token-level optimization pressure during supervised training. However, our design differs from DFT in two key aspects. First, DFT determines the loss weight according to the model's predicted probability of each target token, whereas our Tag-Loss assigns weights according to the structural and semantic role of each token in the generated response. Second, DFT mainly aims to stabilize token-level gradients for general reasoning, while our objective explicitly emphasizes task-critical spans for multimodal forgery detection, including spatial grounding coordinates and the final fake/real prediction.

Therefore, Tag-Loss can be viewed as a task-aware and structure-aware variant of reweighted SFT. It retains the simplicity and stability of supervised fine-tuning, while injecting domain-specific inductive bias into the optimization objective. In our setting, this is particularly important because different tagged spans contribute unequally to the final evaluation: reasoning tokens provide interpretability, bounding-box tokens determine fine-grained localization quality, and answer tokens directly decide image-level detection accuracy.

E Ablation Study of Tag-loss

Table 8: Ablation study of tag-level loss weights.

w_{think}	w_{bbox}	w_{answer}	LOKI	Chameleon
1.0	1.0	1.0	62.15	72.84
1.0	1.5	2.0	68.40	74.50
1.0	2.0	2.0	67.33	73.68
1.0	2.5	2.0	66.50	74.92

Table 8 studies the influence of different tag-level loss weights. Compared with standard SFT, where all tagged spans are assigned an equal weight, introducing larger weights for <bbox_2d> and <answer> consistently improves the performance on both OOD benchmarks. Specifically, the mild setting with $(w_{\text{think}}, w_{\text{bbox}}, w_{\text{answer}}) = (1.0, 1.5, 2.0)$ achieves the best result on LOKI, improving the accuracy from 62.15 to 68.40, with a gain of $+6.25$ points. This indicates that moderately emphasizing spatial grounding and final prediction tokens can effectively strengthen task-critical supervision and improve generalization beyond standard SFT.

We further observe that increasing the weight of <bbox_2d> does not monotonically improve performance. When w_{bbox} is increased from 1.5 to 2.0 or 2.5 while keeping $w_{\text{answer}} = 2.0$, the perfor-

You are a visual reasoning assistant.
Given a image summary and an input image, identify the image region that contains information useful for answering the question.
Describe the given region and explicitly include the key visible text/numbers from that region.
Return JSON only:

```
{
  "bbox_2d": [x1, y1, x2, y2],
  "area_description": "Describe the given region, including
  its spatial context and all question-relevant visible
  text/numbers exactly as shown in the image."
}
```

Requirements:

1. Include exact visible text/numbers in `area_description`.
2. Include labels, row/column names, cell values, axis values, or legends if relevant.
3. List multiple relevant values explicitly.
4. Do not hallucinate missing text.

Input:
Image Summary: {region_summary}
Image: {image}

Figure 5: Prompt for region description and text extraction.

mance on LOKI gradually drops from 68.40 to 67.33 and 66.50, suggesting that overly emphasizing spatial coordinates may over-constrain the model toward local grounding signals and weaken its holistic forgery reasoning ability. Nevertheless, the setting $(1.0, 2.5, 2.0)$ obtains the best performance on Chameleon, reaching 74.92, which implies that stronger localization supervision can still be beneficial for benchmarks that require more fine-grained visual discrepancy perception.

F R2F Data Generation Pipeline

The goal of our R2F data generation pipeline is to construct high-quality cold-start and reinforcement-learning data for teaching the model to perform region-aware forensic reasoning. Unlike standard image-level supervision that only provides a binary label, our pipeline aims to distill a structured reasoning process that explicitly connects visual evidence, localized regions, and the final judgment. Specifically, we build the data from the training subsets of existing public benchmarks, and convert each sample into a tag-structured trajectory containing concise reasoning, region grounding, and task-specific answers. This design encourages the model to learn the three-stage output format and reduces its sensitivity to prompt variations during inference.

Bbox-grounded CoT Generation. For each training sample, we first use the available annotation or pseudo-annotation to obtain suspicious regions associated with the image-level label. These regions are represented as bounding boxes and serve as explicit visual anchors for reasoning. Based on the global image, the localized regions, and the corresponding ground-truth label, we prompt the teacher model to generate a concise chain-of-thought trajectory. The generated response follows a unified tag-based format. To avoid overlong or noisy explanations, we require the teacher model to focus only on discriminative visual evidence, such as unnatural texture, inconsistent boundaries, local artifacts, semantic conflicts, or abnormal generation patterns. This produces compact and instruction-following trajectories that are suitable for cold-start supervised fine-tuning. Follow by ZWZ (Wei et al., 2026), the prompt can be seen in Figure 5.

IoU-based Data Filtering. Although teacher-generated reasoning can provide useful supervision, the generated bounding boxes may still contain noisy or weakly grounded regions. Therefore, we further apply an IoU-based filtering strategy to ensure the spatial reliability of the generated trajectories. Specifically, for samples with available localization annotations, we compute the IoU between the generated bounding box and the reference region. Samples whose IoU falls below a predefined threshold are removed from the cold-start set. This filtering step ensures that the retained trajectories are not only label-consistent but also spatially grounded. For samples without reliable region annotations, we retain only those whose generated answer is consistent with the original label and whose reasoning follows the required tag format.

Trajectory Construction for PA-RLVR. After obtaining the cold-start data, we further construct multi-turn trajectory samples for the subsequent PA-RLVR stage. Different from the cold-start SFT data, which mainly teaches the model the basic output format and reasoning style, the PA-RLVR trajectories are designed to support reward-based optimization over the reasoning process. Each trajectory contains multiple reasoning and grounding steps, allowing the model to learn how to progressively inspect suspicious regions and refine its decision. This stage encourages the model to produce more accurate, stable, and verifiable region-aware reasoning under different prompts and visual inputs.

Finally, our pipeline produces 40,935 high-quality instruction-tuning samples for the cold-start SFT phase and 66,324 multi-turn trajectory samples for the subsequent PA-RLVR stage. These samples provide complementary supervision: the former establishes the basic structured reasoning ability, while the latter further improves the robustness and accuracy of the model through process-level optimization.

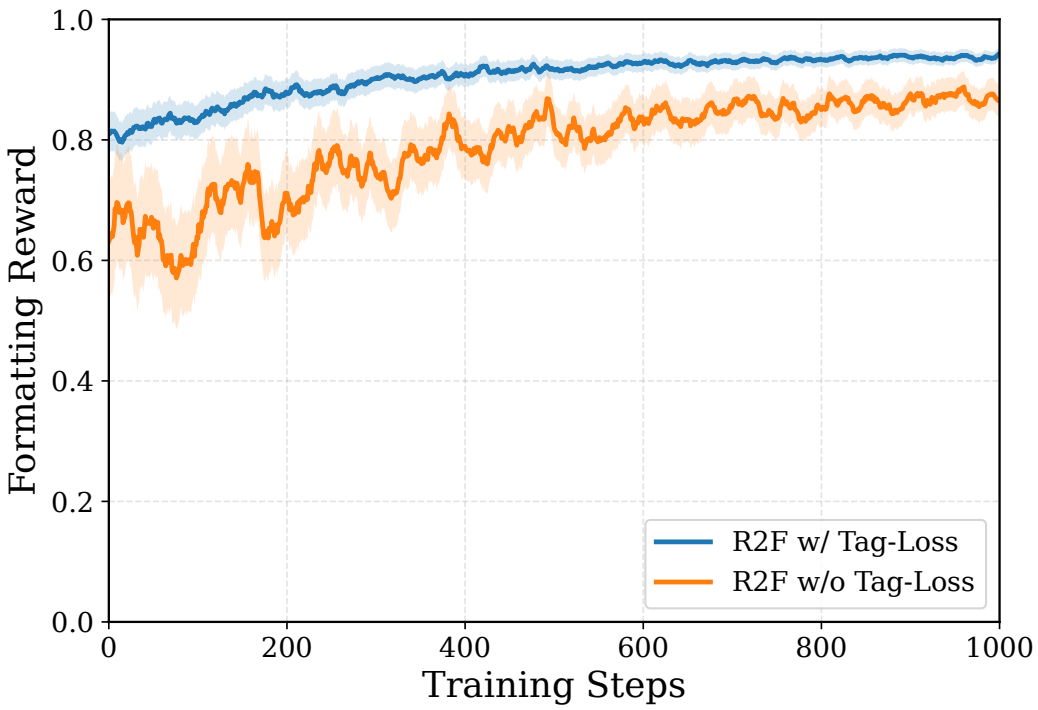

Figure 6: Effect of Tag-Loss on RL Training Stability.

G Effect of Tag-Loss on RL Training Stability.

We introduce Tag-Loss during the cold-start SFT stage to strengthen the model's awareness of structured tag generation. This design is particularly important for PA-RLVR, where stable multi-turn reasoning and reward computation rely on well-formatted intermediate outputs. As shown in Fig-

ure 6, R2F with Tag-Loss starts with a high formatting reward of around 0.8 and maintains a smooth upward trend throughout training. In contrast, removing Tag-Loss results in a much lower initial formatting reward, around 0.6, accompanied by substantial reward oscillations in the early and middle stages. Although the model without Tag-Loss eventually becomes more stable, its convergence is slower and its training process is more sensitive to optimization instability. In practice, this setting also restricts the use of larger learning rates, since aggressive updates can easily lead to entropy collapse. These results indicate that Tag-Loss provides a better cold-start initialization for PA-RLVR and improves the stability of subsequent reinforcement learning.

Table 9: Sensitivity analysis of prompt variants on the LOKI evaluation benchmark (Accuracy). The enhanced prompt improves detection performance on the Person, Medicine, and Doc categories.

Model	**Scene**	**Animal**	**Person**	**Object**	**Medicine**	**Doc**	**Satellite**	**Overall**
Random Choice	21.6	18.3	18.6	26.0	22.2	22.1	-	18.0
Human	24.0	25.8	19.9	26.9	26.1	22.1	-	27.3
AIDE (ICLR'25)	-	89.9	62.5	96.5	53.4	49.7	39.3	63.1
ForensicMOE (ICCV'25)	80.9	81.0	61.3	98.8	69.2	47.6	86.9	74.4
Claude-3.5-Sonnet	51.6	51.6	55.2	51.4	51.9	59.1	50.9	53.6
GPT-4o	70.1	69.7	84.4	70.3	54.3	60.1	45.0	63.4
Qwen3-VL-30B-A3B-Instruct	52.7	32.7	58.1	24.6	50.0	49.6	33.6	42.8
ThinkFake	78.5	87.2	71.7	89.4	64.5	54.2	82.2	75.4
FakeVLM (NeurIPS'25)	98.9	89.7	73.8	94.1	69.3	85.0	97.2	84.3
FakeShield (ICLR'25)	50.0	78.1	58.5	93.7	49.8	45.2	66.7	62.6
Veritas (ICLR'26 oral)	71.1	74.4	78.3	66.5	62.5	63.1	69.2	68.9
Ivy-xDetector	88.7	84.3	70.6	89.2	55.1	49.6	70.4	71.2
Ours (R2F)	**95.4**	**95.5**	81.2	92.2	**77.3**	**92.3**	**96.6**	**90.1**

Table 10: Localization performance comparison on CASIA1+ (Dong et al., 2013), AutoSplice (Jia et al., 2023), RTM (Luo et al., 2025), and OSTF (Qu et al., 2025) benchmarks. We report image-level AUC and localization IoU, where higher values indicate better detection and localization performance. Missing entries indicate that the corresponding method does not provide valid localization outputs or the metric is not reported.

Model	**CASIA1+**		**AutoSplice**		**RTM**		**OSTF**	
	AUC	**IoU**	**AUC**	**IoU**	**AUC**	**IoU**	**AUC**	**IoU**
DTD (Qu et al., 2023)	0.468	-	0.429	-	0.451	0.172	0.603	0.563
TruFor (Guillaro et al., 2023)	0.940	0.692	0.646	<u>0.393</u>	0.599	-	0.594	-
Mesorch (Zhu et al., 2025)	<u>0.950</u>	**0.840**	<u>0.676</u>	0.249	0.472	-	0.512	-
Ours (R2F)	**0.961**	<u>0.833</u>	**0.701**	**0.443**	**0.580**	**0.492**	**0.732**	**0.659**

H LOCALIZATION PERFORMANCE COMPARISON.

Table 10 compares R2F with representative image manipulation localization methods across CASIA1+, AutoSplice, RTM, and OSTF. Overall, R2F achieves the best AUC on all four benchmarks, demonstrating strong image-level manipulation detection capability across diverse localization scenarios. In terms of region-level localization, R2F also obtains the best IoU on AutoSplice, RTM, and OSTF, and remains highly competitive on CASIA1+, where it is only slightly lower than Mesorch.

I PROMPT SENSITIVITY ANALYSIS

As shown in Table 9, we conduct a prompt-sensitivity study on the LOKI benchmark. R2F strengthens the detection performance on the Person, Medicine, and Doc categories, while maintaining comparable performance on the remaining categories.

Table 11: Effect of training data and training paradigm. We compare standard SFT with different data configurations against our full R2F pipeline. Cold-start SFT uses only the 40K cold-start distillation data. SFT w/ RL Data directly treats the 66K PA-RLVR trajectory data as supervised fine-tuning data. SFT w/ All Data performs supervised fine-tuning on the union of the 40K cold-start data and 66K PA-RLVR data. Ours first performs cold-start SFT on the 40K data and then applies PA-RLVR on the 66K trajectory data.

Method	SFT Data	RL Data	LOKI	Chameleon	OpenMMSec				
					DeepFake	AIGC	IMDL	Doc	Avg.
Cold-start SFT	$R2F_{SFT}$-40K	–	62.15	72.84	74.36	72.91	70.48	68.27	71.51
SFT w/ RL Data	$R2F_{PA\text{-}RLVR}$-66K	–	66.82	73.46	76.18	74.63	72.35	69.84	73.25
SFT w/ All Data	$R2F_{SFT}$-40K + $R2F_{PA\text{-}RLVR}$-66K	–	72.54	75.38	78.92	77.64	75.31	72.86	76.18
Ours	$R2F_{SFT}$-40K	$R2F_{PA\text{-}RLVR}$-66K	90.14	79.02	84.26	83.95	82.14	79.98	82.58

J Information About Use of AI Assistants

We used AI assistants only for language polishing, grammar correction, and translation during the writing process. All scientific ideas, methodological designs, experiments, analyses, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of this paper.

K The Effect of Training Data

Table 11 studies the effect of different data usage strategies and training paradigms. Cold-start SFT already provides a reasonable initialization, achieving 62.15% and 72.84% accuracy on LOKI and Chameleon, respectively. Directly treating the PA-RLVR trajectories as standard supervised data brings only moderate improvements, suggesting that these trajectories alone are insufficient to fully activate fine-grained perceptual reasoning. Combining the cold-start and PA-RLVR data under pure SFT further improves performance, but the gains remain limited compared with our full pipeline. In contrast, our method substantially improves out-of-domain generalization, especially on LOKI, where the accuracy increases from 62.15% to 90.14%. It also achieves the best average performance on OpenMMSec, demonstrating that PA-RLVR is not merely an additional data source but an effective optimization paradigm for aligning the model with verifiable perceptual cues.