MMSEC-R1: TEST-TIME SCALING FOR FAKE IMAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Current Multimodal Large Language Models (MLLMs) typically treat image forgery detection and localization as separate or loosely coupled tasks, addressing them through isolated detection and localization stages. In this paper, we redefine this paradigm by proposing Fake Image Reasoning (FIR), a unified paradigm that equates authenticity verification with the rigorous reasoning of forgery traces-essentially, an image is fake only if specific manipulated regions can be reasoned out. We introduce MMSec-R1, the first training framework to explore Test-Time Scaling (TTS) for FIR via a multi-stage unified Reinforcement Fine-Tuning (RFT) strategy. Unlike traditional methods, MMSec-R1 treats precise localization as the definitive correctness verification, compelling the model to self-evolve thinking patterns from limited cold-start data. We validate our approach by training models across different scales (4B and 8B), both achieving superior performance. Our results demonstrate that scaling inference-time compute via RL effectively induces complex forensic reasoning in compact models, bridging the gap between detection confidence and visual evidence.

1 INTRODUCTION

The proliferation of advanced generative models (Wang et al., 2025c; Team, 2025; Zhang et al., 2024) has precipitated a crisis in digital trust. While these advancements foster creativity, they simultaneously pose severe security risks, facilitating the proliferation of disinformation (Diel et al., 2025), and malicious forgeries (Clark & Lewandowsky, 2026). Consequently, establishing reliable mechanisms has become an imperative challenge in the field of fake image detection and localization.

Existing methods for this problem predominantly fall into two categories: discriminative deepfake detectors (Tan et al., 2026c; Yan et al., 2025; Zhong et al., 2024) and Multimodal Large Language Models (MLLMs) (Wen et al., 2025; Xu et al., 2025). Traditional detectors typically operate as "black boxes," treating forgery detection as a binary classification task (Yan et al., 2024). While effective on specific benchmarks, they output opaque confidence scores without providing interpretability, leaving users unable to verify *why* an image is flagged as fake (Wen et al., 2025). Recently, MLLMs have been adapted to address this by generating textual explanations. However, a critical limitation remains: current methods often decouple the **judgment** (detection score) from the **evidence** (visual localization). An MLLM might correctly classify an image as "fake" based on statistical artifacts but hallucinate a textual explanation unrelated to the actual manipulated regions (Ye et al., 2025). This disconnection undermines trust and limits the practical utility of forensic tools in sensitive scenarios. In addition, by aligning visual features with linguistic semantics, MLLMs can theoretically articulate forensic evidence. However, the true potential of MLLMs for intrinsic forensic reasoning **remains largely unexplored**. Current SOTA approaches (Wen et al., 2025; Li et al., 2025a) predominantly treat FIR as a pattern detection task (System 1), relying heavily on SFT with human-annotated Chain-of-Thought (CoT) data. This reliance on Cold Start supervision limits the model's capabilities to merely mimicking human-written styles without grasping the substance of visual scrutiny. Consequently, these models often suffer from hallucinations, where textual explanations are decoupled from visual reality, and require cascading external tools like SAM (Ravi et al., 2024) to fix spatial misalignment (Xu et al., 2025). The question remains: *Can rigorous forensic reasoning emerge autonomously from MLLMs without explicit human demonstration?*

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.

To bridge this gap, we argue that robust forgery detection requires a paradigm shift from simple classification to **Fake Image Reasoning (FIR)**. We posit that the authenticity of an image is not merely a label but a logical conclusion derived from visual evidence. Under the FIR paradigm, detection and localization are intrinsically unified: an image is determined to be "fake" *if and only if* verifiable forgery traces can be localized and reasoned out; otherwise, it is deemed "real". This reformulation transforms the task from pattern recognition to a complex reasoning problem, demanding models that can perform region-level scrutiny and logical deduction.

In this paper, we introduce **MMSec-R1**, the first MLLM framework designed to master FIR through **Test-Time Scaling (TTS)**. Inspired by recent advances in reasoning models (OpenAI et al., 2024), we hypothesize that forensic capabilities can emerge from scaling inference-time computation rather than solely scaling model parameters. To achieve this, we propose a multi-stage RFT strategy utilizing Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) Yu et al. (2025). Unlike traditional supervised learning that relies on dense, expensive mask annotations, our approach utilizes a "cold start" with limited data followed by an RFT phase where the model is rewarded for precise localization. This mechanism forces the model to self-evolve thinking patterns Li et al. (2025b), deliberately slowing down to scrutinize visual details, thereby aligning its final verdict with verifiable spatial constraints.

We validate our approach by training MMSec-R1 variants across efficient model scales (MMSec-R1-4B and MMSec-R1-8B). Surprisingly, our results show that even compact models, when equipped with TTS and RFT, can achieve state-of-the-art (SOTA) performance, outperforming significantly larger counterparts and expert MLLMs. This suggests that the bottleneck in current image forensics is not parameter count, but the lack of structured reasoning.

Our contributions can be summarized as follows:

- We formally define the **Fake Image Reasoning (FIR)** paradigm, unifying forgery detection and localization into a single, evidence-based reasoning process.
- We propose **MMSec-R1**, the first framework to explore TTS for image forensics. We introduce a unified RFT training strategy that enables models to self-evolve visual scrutiny capabilities from minimal supervision.
- We demonstrate that our approach yields superior performance on both detection and localization benchmarks using efficient 4B and 8B models, validating that verifiable reasoning is key to scalable and interpretable image security.

2 Related Work

2.1 Fake Image Reasoning

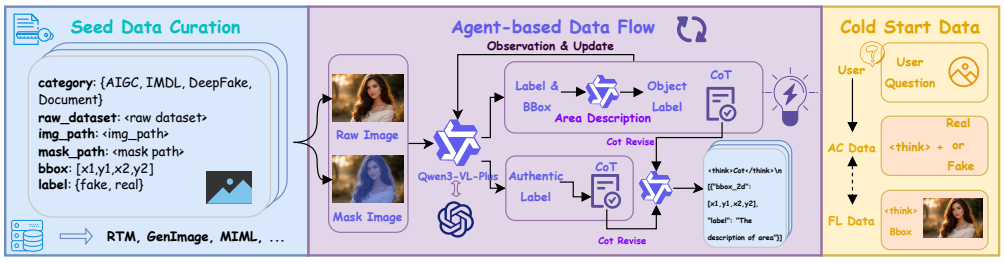

Figure 1: Pipeline for cold-start data construction. We adopt a label-first and mask-based data selection strategy to construct cold-start data by synthesizing raw labels with metadata through LLM prompting. This process transforms sparse supervision into dense, natural language reasoning chains, providing the necessary "warm-up" for subsequent reinforcement learning.

With the advent of foundational vision-language models (VLMs) like CLIP (Radford et al., 2021), recent research has shifted towards leveraging semantic inconsistencies for detection. Methods such as (Yan et al., 2025; 2024; Tan et al., 2026c) utilize pre-trained features to improve generalization

across different generative domains. However, these discriminative approaches still lack the capability to provide **user-friendly reasoning** or precise localization of manipulated regions, limiting their utility in real-world forensic scenarios where interpretability is paramount.

To address the opacity of traditional detectors, MLLMs have been introduced to generate natural language explanations alongside detection decisions. Recent frameworks Wen et al. (2025); Keita et al. (2024); Jiang et al. (2025); Tan et al. (2026b) fine-tune LLaVA-based architectures Liu et al. (2023) or Qwen2.5-VL (Bai et al., 2025b) on forensic instruction datasets to describe forgery traces. Despite their promise, existing methods face critical challenges. First, they rely heavily on SFT with human-annotated reasoning chains, which is expensive and unscalable for subtle forgery types. Second, MLLMs often suffer from *hallucination* and *decoupling*, where generated explanations fail to align with visual evidence.

2.2 TTS AND RL

The emergence of reasoning models (Li et al., 2025b), exemplified by OpenAI's o1 and DeepSeek-R1 (Guo et al., 2025), has revolutionized LLM capabilities by scaling test-time compute. Unlike standard System 1 rapid response models, these approaches utilize RL to encourage LLMs to think before answering, utilizing internal monologues to break down complex problems. A key driver of this success is RFT, where ground-truth outcomes serve as rigorous signals to optimize the reasoning policy without a critic model (e.g., via GRPO (Shao et al., 2024)). However, applying RFT to the visual domain is non-trivial due to the lack of clear verifiable intermediate steps in general vision tasks. **MMSec-R1** bridges this gap by identifying *Forgery Localization* as a definitive verifier for visual reasoning. We propose the first framework to stimulate forensic reasoning capabilities through pure RL, demonstrating that spatial constraints (ground-truth bounding boxes) are sufficient to induce sophisticated visual scrutiny and self-correction patterns without human demonstration.

3 METHODOLOGY

3.1 COLD-START DATA CONSTRUCTION

As shown in Figure 1, Our training pipeline commences with the construction of a high-quality seed dataset derived from OpenMMSec (Tianchi, 2025). Unlike previous approaches that rely on massive but noisy datasets, we focus on reasoning density and quality to establish a strong "System 1" prior. We implement a three-stage curation strategy to generate reliable CoT data: (1) **Uniform Source Sampling.** To ensure domain diversity without overwhelming the model with redundant samples, we avoid using the entire OpenMMSec training set directly. Instead, we perform uniform sampling across the various data sources within OpenMMSec. This yields a balanced collection of seed images covering diverse forgery types and distinct visual domains, laying a generalized foundation for subsequent reasoning tasks; (2) **Strong-to-weak Distillation.** Since existing forensic datasets (Qu et al., 2025; Luo et al., 2024) typically provide only binary labels or masks without explanatory logic, we employ `Qwen-VL-Plus` as a teacher model to hallucinate valid reasoning paths. Using the specific prompt provided in Appendix F, we guide the teacher model to generate step-by-step CoT explanations that logically connect visual artifacts to the final detection and localization results; (3) **Artifact Removal and Refinement.** Raw outputs from the teacher model often contain conversational fillers or rigid template markers (e.g., "The user label indicates...") that disrupt the flow of internal reasoning. To address this, we apply a rule-based filtering and rewriting pipeline. We strip these format artifacts

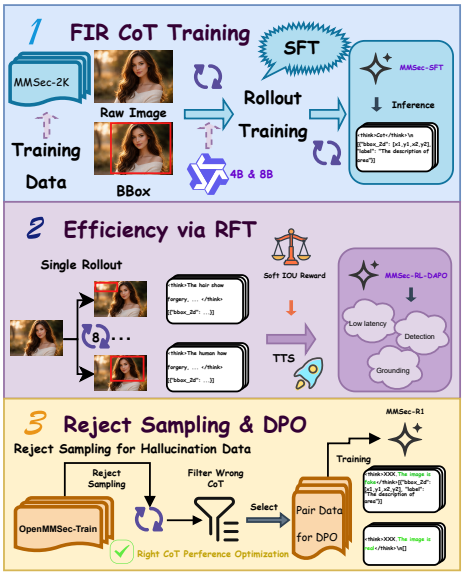

Figure 2: The training framework of MMSec-R1.

and restructure the text into a coherent, first-person reasoning stream, ensuring the data aligns strictly with the intended CoT format; (4) After applying this rigorous pipeline, we curated a final set of fewer than 3,000 high-quality samples. This compact dataset serves as the cold-start foundation, enabling MMSec-R1 to learn basic forensic concepts before scaling up capabilities via RL. The detailed prompt used in our data distillation pipeline is provided in Appendix F.

3.2 TASK-AWARE DAPO

Standard RL typically relies on a Critic model for value estimation, which doubles memory overhead and complicates training. GRPO Shao et al. (2024) eliminates the Critic by employing group-based relative rewards as a baseline. Formally, for a given image-query pair q, we sample a group of outputs $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. The advantage A_i for the i-th output is computed by normalizing rewards within the group:

$$A_i = \frac{R(o_i) - \text{mean}(\{R(o_j)\}_{j=1}^{G})}{\text{std}(\{R(o_j)\}_{j=1}^{G}) + \epsilon}, \tag{1}$$

where $R(o_i)$ denotes the verifiable reward. While GRPO is efficient, it often suffers from training instability and gradient vanishing in long-chain forensic reasoning. To address this, we adopt DAPO Yu et al. (2025), which normalizes the policy gradient loss at the token level to ensure every reasoning step contributes equally. The objective function is defined as:

$$J_{\text{DAPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{t=1}^{L_i}\min\Big(\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon_l, 1+\epsilon_h)A_i\Big)\right], \tag{2}$$

where L_i is the length of output o_i, and the importance ratio is $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$.

Overlong Reward Shaping. To prevent the model from exploiting the reward mechanism via infinitely long, nonsensical reasoning chains, DAPO introduces a soft penalty for outputs exceeding a length threshold L_{max}. A linear penalty is applied to $R(o_i)$ if the length $L_i > L_{soft}$, guiding the model toward conciseness without hard truncation.

By integrating these components, MMSec-R1 effectively explores the vast reasoning space of image forensics, stabilizing the emergence of complex scrutiny behaviors.

3.3 VERIFIABLE REWARD DESIGN

To substitute explicit reasoning supervision, we design a rigorous reward mechanism where Forgery Localization serves as the primary verifier.

Format Reward (R_{fmt}). We strictly enforce format integrity to enable parsing. The output must follow: $Y = \texttt{<think>}T_{\text{reason}}\texttt{</think>}\{\texttt{"bbox_2d":}\ldots\}$. We assign $R_{fmt} = 1.0$ only if the output contains valid tags and parses into a valid JSON with required keys; otherwise $R_{fmt} = 0.0$.

Task-Aware Reward (R_{task}). We define task-aware rewards based on the data type: (1) **Tampering Localization**: For edited images with ground truth boxes G_{gt}, we use Soft IoU. To fit the metric into the optimization target, we define:

$$\text{SIoU} = \frac{2\sum_i \max_j \text{IoU}(g_i, \hat{g}_j) \cdot \sum_j \max_i \text{IoU}(\hat{g}_j, g_i)}{\sum_i \max_j \text{IoU}(g_i, \hat{g}_j) + \sum_j \max_i \text{IoU}(\hat{g}_j, g_i) + \epsilon} \tag{3}$$

where $g \in G_{gt}$ and $\hat{g} \in \hat{G}_{pred}$. For pristine images ($G_{gt} = \emptyset$), $R_{task} = 1.0$ if $\hat{G}_{pred} = \emptyset$, else 0. (2) **Authentic Detection**: For synthetic images, $R_{task} = 1.0$ if the binary label is correct. The calculation process of Soft IoU is provided in Appendix C.

Total Reward. The final reward is a weighted sum: $R = \lambda_{task} R_{task} + \lambda_{fmt} R_{fmt}$.

3.4 Forgery Trace-Prediction Consistency Alignment via DPO

While DAPO effectively boosts the FIR via outcome-based rewards, it primarily optimizes the final prediction Y, potentially neglecting the semantic consistency of the intermediate reasoning chain T_{reason}. We observe a "schizophrenic" phenomenon in the RFT model: the `<think>` block may conclude an image is authentic, yet the subsequent JSON output predicts tampering artifacts to maximize the reward. To eliminate such reasoning-prediction inconsistencies and align the model's internal scrutiny with its output, we introduce a Direct Preference Optimization (DPO) (Rafailov et al., 2023) stage.

Preference Pair Construction. We employ a rejection sampling strategy to curate a high-quality preference dataset $\mathcal{D}_p$. Specifically, for a given prompt x, we sample $K = 8$ distinct responses $\{o_1, \ldots, o_K\}$ from the DAPO-converged policy π_{RL}. We then employ a powerful judge model, Qwen3-Max-2025-09-23 (Yang et al., 2025), to automatically annotate these responses based on two criteria: (1) **Accuracy**: Whether the final prediction matches the ground truth; (2) **Consistency**: Whether the reasoning trace T_{reason} logically supports the final JSON output.

Based on the judge's feedback, we construct preference pairs (x, y_w, y_l):

- **Chosen (y_w)**: Responses that are both accurate and logically consistent.
- **Rejected (y_l)**: Responses that are accurate in prediction but exhibit *hallucinated reasoning* (e.g., describing an image as pristine in the chain-of-thought but outputting forgery bounding boxes), or responses that are simply incorrect.

Through this process, we filter and construct a compact alignment dataset of 1k pairs, focusing on correcting the most severe reasoning detachments. The Prompt can be seen in Appendix 17.

Optimization Objective. We treat the DAPO-trained model as the reference policy π_{ref} and optimize the policy π_θ to maximize the margin between consistent and inconsistent reasoning paths. Defining the implicit reward ratio as $h_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, the objective is simplified as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}_p} \left[\log \sigma \left(h_\theta(x, y_w) - h_\theta(x, y_l) \right) \right] \tag{4}$$

where σ is the sigmoid function (Zhai et al., 2023) and β is a hyperparameter controlling the deviation from the reference policy. By minimizing $\mathcal{L}_{\text{DPO}}$, MMSec-R1 learns to suppress hallucinated reasoning traces, ensuring that the explicit CoT strictly causes the final forensic verdict.

4 Experiments

4.1 Experimental Setup

Baselines. To comprehensively evaluate the capabilities of MMSec-R1, we compare it against a diverse set of MLLMs across three categories: (1) **General MLLMs**: We select representative open-source models known for their general visual understanding, including InternVL3.5 (Wang et al., 2025b), LLaVA-OneVision-1.5 (An et al., 2025), MiMo-VL-7B-RL-2508 (Xiaomi, 2025) and Qwen3-VL-thinking (Bai et al., 2025a). (2) **Reasoning MLLMs**: We include recent models capable of advanced visual reasoning chains, such as M2-Reasoning (AI et al., 2025), MiniCPM-V 4.5 (Yao et al., 2024), Visionary-R1 (Xia et al., 2025), and VL-Rethinker (Wang et al., 2025a). (3) **Expert MLLMs**: We compare against domain-specific forensic models, specifically FakeVLM (Wen et al., 2025), and our proposed MMSec-R1.

Datasets. We utilize the **OpenMMSec** benchmark (Tianchi, 2025), a comprehensive dataset designed for fake image reasoning. OpenMMSec encompasses four distinct expert categories: DeepFake (face manipulation), IMDL (image manipulation detection and localization), AIGC (fully

Table 1: Results on In-domain (ID) performance on FIR tasks. For authentic classification, ACC and F1 is reported; for forgery localization (IMDL and Document), Soft IoU is reported; * denotes the model **lacks grounding capability** and does not compute Soft IoU.

Method	Size	OpenMMSec In-domain Evaluation				
		Authentic classification		forgery localization		Average
		Deepfake	AIGC	IMDL	Document	
General MLLMs						
InternVL3.5	4B	52.63/51.88	52.64/54.96	55.02/54.92/1.16	45.18/53.73/0.95	52.34/52.39/1.15
	8B	50.64/50.59	52.49/54.87	52.65/52.23/0.09	60.78/67.79/0.00	53.53/53.52/0.09
LLaVA-OneVision-1.5	4B	53.72/53.14	47.14/49.73	57.19/53.22/15.81	29.65/33.39/3.63	49.94/47.68/15.08
	8B	56.05/49.18	60.13/58.90	57.63/51.28/2.30	21.84/23.09/0.08	51.49/45.03/2.17
Qwen3-VL	4B	58.88/56.59	63.40/64.13	60.06/55.83/27.25	44.04/51.71/10.28	57.56/55.09/26.24
	8B	57.44/53.43	65.26/66.04	61.73/59.24/26.65	53.40/61.30/5.52	59.82/58.10/25.39
MiMo-VL	7B	60.62/56.96	72.53/72.07	58.16/52.14/28.75	20.35/20.14/7.20	54.62/49.34/27.47
Reasoning MLLMs						
MiniCPM-V 4.5	8.7B	58.62/52.22	58.45/57.55	68.93/67.61/26.31	79.51/81.99/1.69	66.00/63.56/25.16
Visionary-R1*	3B	56.99/54.67	68.82/67.24	56.22/47.65/-	14.11/7.31/-	51.23/44.44/-
VL-Rethinker*	7B	54.22/42.94	65.78/62.56	58.10/51.41/-	19.43/17.33/-	51.75/43.31/-
M2-Reasoning*	7B	56.85/55.55	58.13/60.07	63.16/60.43/-	22.06/22.60/-	53.70/52.76/-
Expert MLLMs						
FakeVLM*	7B	78.87/78.79	88.13/88.35	62.99/58.83/-	11.28/2.62/-	62.10/60.13/-
MMSec-R1	4B	79.11/80.90	95.98/96.09	94.34/92.90/55.20	82.01/81.66/20.01	87.86/87.89/37.61
	8B	83.28/82.20	96.07/97.17	96.98/97.58/62.23	85.03/85.23/31.32	90.34/90.55/46.78

Table 2: Results on Out-of-Domain (OOD) performance on FIR tasks. For authentic classification, ACC and F1 is reported; for forgery localization (IMDL and Document), Soft IoU is reported; * denotes the model **lacks grounding capability** and does not compute Soft IoU.

Method	Size	OpenMMSec Out-of-domain Evaluation				
		Authentic classification		forgery localization		Average
		Deepfake	AIGC	IMDL	Document	
General MLLMs						
InternVL3.5	4B	40.04/39.59	49.95/46.77	49.51/46.36/0.37	30.34/34.40/0.07	44.11/41.57/0.28
	8B	39.63/38.91	50.59/47.11	49.35/45.30/0.01	31.77/36.04/0.00	44.34/41.42/0.01
LLaVA-OneVision-1.5	4B	48.45/50.16	55.57/55.64	56.66/52.36/16.61	30.18/32.63/2.04	50.27/49.81/12.43
	8B	37.55/31.63	56.32/54.85	57.04/51.32/3.54	25.21/25.71/0.16	46.88/42.73/2.57
Qwen3-VL	4B	41.59/40.38	64.50/64.04	59.01/54.37/25.35	49.13/55.58/1.92	54.13/52.74/18.64
	8B	39.69/36.77	66.44/66.18	60.33/57.09/23.88	51.40/58.04/1.92	54.85/53.69/17.58
MiMo-VL	7B	42.19/39.76	67.87/67.44	58.17/52.61/32.91	32.46/35.82/2.63	53.11/50.84/24.22
Reasoning MLLMs						
MiniCPM-V 4.5	8.7B	66.90/62.83	71.86/70.03	67.90/67.09/23.08	71.68/75.60/1.44	69.32/67.83/17.98
Visionary-R1*	3B	45.13/44.98	62.03/60.52	57.09/50.00/-	18.73/13.18/-	49.85/46.39/-
VL-Rethinker*	7B	31.82/20.17	56.87/53.75	57.07/50.41/-	24.05/22.44/-	45.30/38.63/-
M2-Reasoning*	7B	61.59/62.46	62.22/62.27	61.17/57.97/-	32.18/35.31/-	57.47/57.96/-
Expert MLLMs						
FakeVLM*	7B	75.78/76.38	86.65/86.57	66.22/63.65/-	13.99/4.62/-	67.79/68.21/-
MMSec-R1	4B	72.76/75.07	87.61/91.98	85.24/84.36/46.04	73.98/73.84/20.99	79.90/81.31/33.52
	8B	74.94/75.02	88.23/90.67	87.61/88.40/56.83	78.03/80.17/22.76	82.20/83.56/39.80

synthetic generation), and Document. The dataset is rigorously split into a training set, an *In-Domain (OpenMMSec-ID)* test set, and an *Out-of-Domain (OpenMMSec-OOD)* test set to evaluate generalization capabilities. Detailed cases of the dataset are provided in Appendix E.

Evaluation Metrics. We employ a dual-metric protocol to assess both detection and localization performance: (1) **Authentic Classification**: We report Accuracy (ACC) and F1-score to evaluate the binary classification performance. (2) **Forgery Localization**: We utilize Soft Intersection over Union (Soft IoU) to measure the precision of the predicted manipulated regions. Note that Soft IoU is calculated exclusively on samples containing ground-truth manipulation masks. To ensure a rigorous and decoupled evaluation, we employ task-specific prompts for classification and localization, respectively. The full list of prompts is available in Appendix F.

4.2 Experimental Results

General MLLMs vs. Expert MLLMs. As shown in Table 1, MMSec-R1 significantly outperforms both General and Reasoning MLLMs, establishing a new SOTA across all metrics. While General MLLMs like InternVL3.5 (Wang et al., 2025b) and Qwen3-VL (Bai et al., 2025a) exhibit robust

Table 3: Test-time compute scaling with different sample budgets. We report average generated tokens, latency, detection performance, and localization performance.

Model / #Samples	Avg Tokens	Latency (s)	ACC	F1	Soft IoU
MMSec-R1-4B					
$N = 1$	810	0.92	89.20	88.75	44.51
$N = 2$	1580	1.75	89.82	89.30	45.12
$N = 4$	3090	3.42	90.41	89.88	45.83
$N = 8$	6020	6.70	90.68	90.15	46.05
MMSec-R1-8B					
$N = 1$	815	1.05	90.34	89.92	46.78
$N = 2$	1605	2.02	90.92	90.48	47.32
$N = 4$	3140	3.95	91.51	91.02	47.96
$N = 8$	6120	7.75	91.83	91.36	48.20

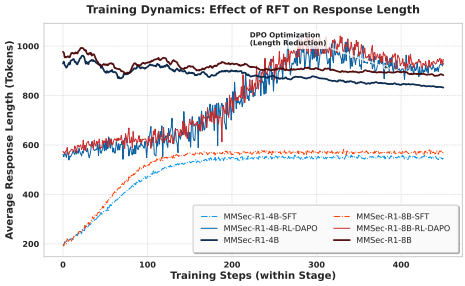

Figure 3: The evolution of response length during training. The curves demonstrate the transition from SFT cold-start to RFT, followed by conciseness optimization in the final stage.

Table 4: Ablation study on in-domain performance of Qwen3-VL-8B (Bai et al., 2025a) on OpenMMSec-ID.

MODEL	COLD START	DAPO	GRPO	DPO	EVALUATION METRIC ACC/SOFT IOU
QWEN3-VL-8B					59.82/25.39
MMSEC-ZERO-8B		✓			78.92/10.41
MMSEC-R1-8B-SFT	✓				52.63/10.76
MMSEC-R1-8B-RL-GRPO	✓		✓		85.80/42.00
MMSEC-R1-8B-RL-DAPO	✓	✓			88.87/43.67
MMSEC-R1-8B	✓	✓	✓	✓	90.34/46.78

semantic understanding, they struggle with high-frequency artifact detection, lagging behind MMSec-R1-8B by an average of 30.52% in ACC. Interestingly, Reasoning MLLMs (e.g., Visionary-R1 (Xia et al., 2025)) show improved performance but still fail to match FIR experts, suggesting that general-purpose MLLMs do not automatically transfer to the nuanced domain of fake image detection. Notably, MMSec-R1 surpasses the previous SOTA expert, FakeVLM (Wen et al., 2025), by substantial margins (e.g., +30.42% F1 on average).

ID vs. OOD. As shown in Table 2, generalization to unseen expert methods is the ultimate test for expert models. On the OOD test set, supervised experts typically suffer significant performance drops due to overfitting to specific ID expert samples. However, MMSec-R1 exhibits remarkable robustness, maintaining high performance on OOD (e.g., retaining about 91% of its ID performance). This suggests that TTS transforms the model from a pattern matcher into a robust visual reasoner, capable of adapting to novel attacks without explicit retraining. A critical observation from our experiments is the correlation between detection accuracy and localization IoU. General MLLMs often achieve moderate detection ACC but low Soft IoU, indicating a tendency to rely on dataset biases or shortcuts rather than true visual evidence. In contrast, MMSec-R1 demonstrates a harmonious improvement in both metrics.

Comparison with SOTA small vision models. We compare MMSec-R1with different SOTA methods in Table 7. We further compare MMSec-R1with state-of-the-art small vision models, including AIDE Yan et al. (2024), Drct Chen et al. (2024), CO-SPY Cheng et al. (2025), Mesorch Zhu et al. (2025), as well as large multimodal forensic models such as Veritas Tan et al. (2026a) and FakeShield Xu et al. (2025) on OpenMMSec. For a fair comparison, all small vision models are retrained on the same training set. Detailed results are provided in Appendix B. We conduct a comprehensive analysis of the training dynamics across different model scales (4B and 8B) to investigate the impact of RFT strategies.

Table 5: Robustness analysis on the OpenMMSec-ID subset under JPEG compression. We evaluate model performance under different JPEG quality factors (QF). The evaluation metrics are Accuracy/Soft IoU.

METHOD	ORIGINAL	95	90	75	50
QWEN3-VL-8B	59.82/25.39	59.77/25.80	59.99/25.55	59.01/24.75	59.07/25.07
MMSEC-R1-8B	90.34/46.78	91.09/46.22	91.00/44.53	90.12/43.29	89.99/42.78

Table 6: Robustness analysis on the OpenMMSec-ID subset under Gaussian blur. We evaluate model performance with increasing blur intensity (σ). The evaluation metrics are Accuracy/Soft IoU.

METHOD	ORIGINAL	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 3.0$	$\sigma = 4.0$
QWEN3-VL-8B	59.82/25.39	58.01/23.60	54.93/23.34	52.79/22.55	51.35/21.48
MMSEC-R1-8B	90.34/46.78	90.20/46.00	91.00/46.34	89.67/43.20	88.61/42.75

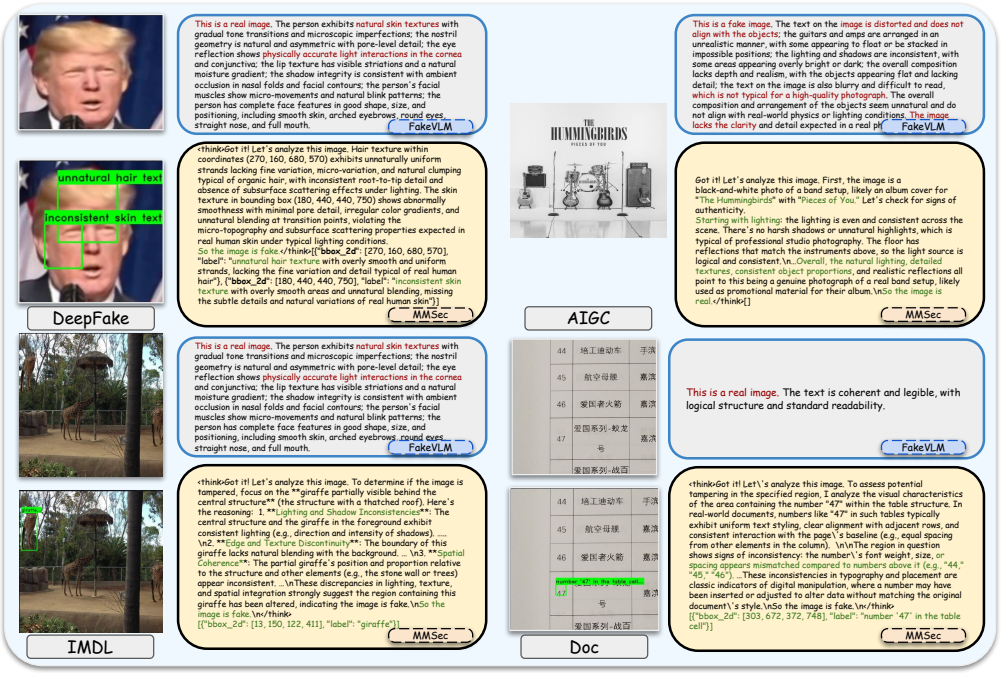

Figure 4: Fake Image Reasoning cases, cover DeepFake, AIGC, IMDL, and Doc (Document). MMSec-R1 outperforms FakeVLM Wen et al. (2025) in fake image detection and localization.

4.3 TEST-TIME SCALING ANALYSIS

As shown in Table 3, both MMSec-R1-4B and MMSec-R1-8B exhibit clear compute-performance scaling. Increasing the sample budget consistently improves image-level detection performance and region-level localization quality. For MMSec-R1-4B, increasing N from 1 to 8 improves ACC from 89.20 to 90.68, F1 from 88.75 to 90.15, and Soft IoU from 44.51 to 46.05. A similar trend is observed for MMSec-R1-8B, where ACC increases from 90.34 to 91.83 and Soft IoU improves from 46.78 to 48.20. These results indicate that the model can exploit additional inference-time computation to produce more reliable forensic judgments and more accurate spatial predictions. Meanwhile, the gains show diminishing returns as the sample budget increases. For example, the improvement from $N = 4$ to $N = 8$ is smaller than that from $N = 1$ to $N = 4$, while the latency and number

Table 7: Performance on Multi Benchmarks (ACC/F1) (GenImage (Zhu et al., 2023), AIGCDetectionBenchmark (Zhong et al., 2023) and Chameleon (Yan et al., 2024)). * denotes the metrics are source from AgentFoX (Yu et al., 2026).

MODEL	VENUE	GENIMAGE	AIGCDETECTIONBENCHMARK	CHAMELEON
DRCT (CHEN ET AL., 2024)	ICML'24	53.53/51.33	47.87/45.76	46.26/43.83
RINE (KOUTLIS & PAPADOPOULOS, 2024)	ECCV'24	51.76/51.38	46.45/46.07	42.43/40.55
SPAI (KARAGEORGIOU ET AL., 2025)	CVPR'25	48.24/50.19	46.10/46.29	48.09/43.48
AIDE (YAN ET AL., 2024)	ICLR'25	86.88/87.42	69.73/70.21	50.89/52.10
CO-SPY (CHENG ET AL., 2025)	CVPR'25	76.30/75.60	72.50/72.00	68.80/68.40
FAKEVLM (WEN ET AL., 2025)	NEURIPS'25	51.96/52.01	54.26/53.89	62.56/67.44
AGENTFOX-8B* (YU ET AL., 2026)	-	72.94/73.36	67.02/65.93	68.39/65.95
AGENTFOX-32B* (YU ET AL., 2026)	-	78.04/78.54	75.71/75.58	70.88/69.14
OURS (MMSEC-R1-4B)	-	**94.04/95.04**	87.08/88.80	73.01/72.67
OURS (MMSEC-R1-8B)	-	93.46/93.42	**89.12/90.00**	**75.22/73.90**

of generated tokens continue to grow approximately linearly. This suggests that moderate TTS budgets already provide a favorable trade-off between accuracy and efficiency. Importantly, since the decoding configuration is fixed across all settings, the observed improvement cannot be attributed to decoding hyperparameter changes. Instead, it provides direct evidence that MMSec-R1benefits from additional test-time reasoning compute.

4.4 QUALITATIVE ANALYSIS

In practical applications, tampered images are significantly more challenging to detect; while FakeVLM (Wen et al., 2025) demonstrates superior performance on AI-generated images, it fails to detect localized image manipulations. Consequently, models must possess the capability to identify local tampering while providing superior interpretability for localization. Figure 4 illustrates the performance of MMSec-R1 on tampered images. This holistic approach significantly enhances the model's ability to detect and localize forgeries in manipulated images. Our analysis reveals that the **Reasoning process** is the primary driver behind the model's final decision accuracy.

4.5 ABLATION STUDIES

The effectiveness of training strategies. As illustrated in Table 4, we conduct a systematic ablation to evaluate the contribution of each training phase. Compared to the base Qwen3-VL-8B, the *MMSec-R1-8B-SFT* (which only utilizes a small amount of CoT data injection) surprisingly shows a slight decline in both ACC and Soft IoU; Our proposed *MMSec-R1-8B-RL* utilizing the DAPO strategy consistently outperforms standard GRPO. By removing the asymmetric clipping bounds (ϵ_h, L_{soft}, and dynamic sampling (Yu et al., 2025). GRPO exhibits less stability in complex FIR tasks. Finally, the integration of **DPO** provides a steady marginal improvement by rectifying hard cases and mitigating the "reward hacking" (Gao et al., 2023) often encountered during the RL phase.

Robustness Analysis. In Table 5 and Table 6, we investigate the robustness of MMSec-R1-8B against common real-world image degradations, specifically JPEG compression and Gaussian blur. Both the backbone Qwen3-VL-8B and our MMSec-R1-8B demonstrate impressive resilience to these perturbations. Although we observe a performance fluctuation of approximately 5% in both ACC and IoU, the model maintains a high degree of reliability, suggesting that the learned reasoning chains are anchored in robust structural artifacts rather than fragile high-frequency noise.

5 CONCLUSION

In this paper, by transitioning from traditional discriminative patterns to a reasoning-driven analytical process, MMSec-R1 demonstrates a superior ability to not only detect sophisticated AIGC forgeries but also provide high-fidelity localization through structured reasoning. Our extensive experiments across 4B and 8B scales reveal that reinforcement learning, specifically through our cold-start SFT, DAPO and DPO strategy, effectively activates the model's thinking capabilities, allowing it to navigate complex forensic scenarios that baffle standard MLLMs.

AI USE STATEMENT

In this work, we used generative AI tools for translation and grammar correction. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for no tasks with recommended disclosure. We have reviewed all AI-assisted work. We manually reviewed all AI-assisted translations and grammar corrections to ensure that they accurately preserved the intended meaning and did not alter the technical content. 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 does not involve human-subject studies or interventions. We use data solely for research purposes and follow applicable requirements regarding data use, privacy, and research integrity. We do not collect or disclose additional personally identifiable or sensitive information beyond what is necessary for the research. We are not aware of any significant ethical risks, conflicts of interest, or legal compliance concerns arising from this work.

REFERENCES

Inclusion AI, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, Yi Yuan, Yifan Mao, Yuting Xiao, and Ziping Ma. M2-reasoning: Empowering mllms with unified general and spatial reasoning, 2025. URL https://arxiv.org/abs/2507.08306.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

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

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

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

Simon Clark and Stephan Lewandowsky. The continued influence of AI-generated deepfake videos despite transparency warnings. *Communications Psychology*, January 2026. ISSN 2731-9121. doi: 10.1038/s44271-025-00381-9. URL https://doi.org/10.1038/s44271-025-00381-9.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Alexander Diel, Tania Lalgi, Finley Sam Mellis, Alexander Teufel, and Alexander Bäuerle. The harm of deepfakes: a scoping review of deepfakes' negative effects on human mind and behavior. *AI & SOCIETY*, December 2025. ISSN 1435-5655. doi: 10.1007/s00146-025-02774-0. URL https://doi.org/10.1007/s00146-025-02774-0.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, et al. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. *arXiv preprint arXiv:2506.00979*, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18706–18717, 2025.

Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, and Abdenour Hadid. Fidavl: Fake image detection and attribution using vision-language model. In *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XXI*, pp. 160–176, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-78304-3. doi: 10.1007/978-3-031-78305-0_11. URL https://doi.org/10.1007/978-3-031-78305-0_11.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, pp. 394–411. Springer, 2024.

Yixuan Li, Yu Tian, Yipo Huang, Wei Lu, Shiqi Wang, Weisi Lin, and Anderson Rocha. Fakescope: Large multimodal expert model for transparent ai-generated image forensics, 2025a. URL https://arxiv.org/abs/2503.24267.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025b. URL https://arxiv.org/abs/2502.17419.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

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*, pp. 110828, 2024.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang,

Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Omni-iml: Towards unified image manipulation localization, 2025. URL https://arxiv.org/abs/2411.14823.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *NeurIPS*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training, 2021. URL https://arxiv.org/abs/2101.06840.

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, 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.

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*, 2026a.

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*, 2026b.

Lei Tan, Shuwei Li, Mohan Kankanhalli, and Robby T. Tan. Aggregating diverse cue experts for ai-generated image detection. In *AAAI*, 2026c.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, 2019.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Tianchi. Open multi-modal ai security evaluation benchmark (openmmsec). https://tianchi.aliyun.com/dataset/210595, 2025. 2025-12-27.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025a. URL https://arxiv.org/abs/2504.08837.

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, 2025b. URL https://arxiv.org/abs/2508.18265.

Yifei Wang, Weimin Bai, colin zhang, Debing Zhang, Weijian Luo, and He Sun. Uni-instruct: One-step diffusion model through unified diffusion divergence instruction. In *NeurIPS*, 2025c. URL https://openreview.net/forum?id=q1acmWAO9r.

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.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14677.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

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*, 2024.

Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. In *ICML*, 2025.

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.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

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.

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.

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, 2026. URL https://arxiv.org/abs/2603.23115.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. doi: 10.1109/TPAMI.2024.3355414.

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.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024. URL https://arxiv.org/abs/2311.12397.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: a million-scale benchmark for detecting ai-generated image. In *NeurIPS*, NIPS '23. Curran Associates Inc., 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.

A LICENSE

We declare that all the original datasets and models used in the experiment are available for research purposes. The license of OpenMMSec can be seen in (Tianchi, 2025).

```
Is this image real or generated by AI? If the image is real, output
an empty JSON. If the image is tampered, locate the tampered
object(s).  For each tampered object, use its specific label
(e.g., text content or general area describe) instead of a fix
''tampered'' label.  Report the object name (label) and the bounding
box coordinates (bbox_2d) in JSON format.
```

Figure 5: Training and evaluation prompt for MMSec-R1.

Table 8: Comparison of State-of-the-Art Methods, including expert models Effort (Yan et al., 2025) and Qwen3-VL (Bai et al., 2025a). All baselines are re-trained on OpenMMSec's training set to ensure a fair comparison. * denotes the results of the Qwen3-VL series after full-parameter SFT on the OpenMMSec training set; **Bold** indicates the best performance; Underline indicates the second-best performance.

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 & Le (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	73.4	80.4
Trufor Guillaro et al. (2023)	CVPR'23	72.2	82.5	80.5	72.3	76.9
Effort Yan et al. (2025)	ICML'25	<u>85.0</u>	81.9	83.7	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
FakeVLM Wen et al. (2025)	NeurIPS'25	75.8	86.6	66.2	14.0	60.7
Veritas Tan et al. (2026a)	ICLR'26	83.4	72.8	66.8	32.1	63.8
FakeShield Xu et al. (2025)	ICLR'25	72.2	73.2	59.5	68.5	68.4
Ivy-xDetector Jiang et al. (2025)	ICMR'26	56.3	83.1	64.2	17.7	55.3
Qwen3-VL-4B-SFT*	-	78.4	84.6	76.8	70.6	77.6
Qwen3-VL-8B-SFT*	-	81.2	86.1	78.4	71.9	79.4
MMSec-R1-4B	-	82.8	<u>96.0</u>	<u>85.8</u>	<u>87.0</u>	<u>87.9</u>
MMSec-R1-8B	-	**88.0**	**96.1**	**89.9**	**87.2**	**90.3**

B Comparison with SOTA small vision models.

As shown in Table 8, we compare MMSec-R1with representative small vision models and recent forensic MLLMs on OpenMMSec. All small vision baselines are retrained on the same OpenMMSec training set for a fair comparison. Traditional vision models achieve competitive results on specific subsets, such as Effort on DeepFake and SegFormer on Doc, but their average performance remains limited due to weak cross-domain generalization. In contrast, forensic MLLMs such as FakeVLM, Veritas, and FakeShield show stronger semantic reasoning ability but still suffer from unstable localization and poor performance on document-style manipulation. Our MMSec-R1consistently achieves the best average performance, reaching 87.9 with the 4B model and 90.3 with the 8B model. Notably, MMSec-R1-8B obtains strong results across all four categories, including 88.0 on DeepFake, 96.1 on AIGC, 89.9 on IMDL, and 87.2 on Doc. These results indicate that the proposed reasoning-oriented training strategy improves both general forgery detection and fine-grained manipulation understanding, while retaining stronger robustness than purely discriminative small vision models.

C Soft IoU Computation

We use Soft IoU as the localization reward for samples with ground-truth forgery regions. Unlike standard IoU, which usually assumes a one-to-one match between predicted and ground-truth boxes, Soft IoU provides a smoother set-level matching score. This is important for FIR because a model may output multiple suspicious regions, while the ground truth may contain one or more manipulated areas. A hard matching rule can overly penalize partially correct predictions, whereas Soft IoU rewards predictions that spatially overlap with the target forgery regions.

Given a set of ground-truth boxes $G = \{g_i\}_{i=1}^{N}$ and predicted boxes $\hat{G} = \{\hat{g}_j\}_{j=1}^{M}$, we compute the best overlap from each ground-truth box to the predicted set and from each predicted box to the ground-truth set. Formally, Soft IoU is defined as:

Algorithm 1 Reward Calculation

Input: Completion C, GT Label y, GT Bbox B_{gt}
Params: $\lambda_{task} \leftarrow 0.9$, $\lambda_{fmt} \leftarrow 0.1$
$T, J \leftarrow$ Parse(C) {Extract think & JSON}
{— 1. Format Check —}
if T is valid **and** J is valid JSON **then**
  $R_{fmt} \leftarrow 1.0$; $B_{pred} \leftarrow J$["bbox_2d"]
else
  return 0.0 {Format failure penalty}
end if
{— 2. Task Verification —}
if B_{gt} exists (Localization Task) **then**
  if $y ==$ "real" **then**
    $R_{task} \leftarrow 1.0$ **if** B_{pred} is empty **else** 0.0
  else
    $R_{task} \leftarrow$ SoftIoU(B_{gt}, B_{pred}) {Equation 3}
  end if
else
  {Detection Task}
  $L_{pred} \leftarrow J$["label"]
  $R_{task} \leftarrow 1.0$ if $L_{pred} = y$ else 0.0
end if
Output: $\lambda_{task} R_{task} + \lambda_{fmt} R_{fmt}$

D Implementation Details

In the Cold Start phase, we initialize the model using the Qwen3-VL backbone (Bai et al., 2025a), where we freeze both the vision encoder and the projector while fully fine-tuning the language model parameters. We train the model for 1 epochs with a learning rate of 1×10^{-5} and a warmup ratio of 0.05, utilizing DeepSpeed ZeRO-3 (Ren et al., 2021) and Flash Attention 2 (Dao, 2023) to optimize memory efficiency. In the subsequent RL phase, we configure the group size $G = 8$, the KL penalty coefficient $\beta = 0.04$, sampling temperature $\tau = 1.05$, and learning rate 1×10^{-5}. Specifically for the DAPO strategy, we employ asymmetric clipping bounds with $\epsilon_h = 0.6$ to encourage exploration, and apply overlong reward shaping with a soft length threshold $L_{soft} = 1024$ to regulate reasoning depth. Finally, for the Forgery Consistency Alignment via DPO, we curate the preference pairs via rejection sampling ($K = 8$) and fine-tune the model for 1 epoch. We adopt a reduced learning rate of 5×10^{-7} with a cosine decay scheduler and set the DPO deviation penalty $\beta_{dpo} = 0.1$ (Rafailov et al., 2023) to maintain stability against the reference policy.

```
Is the image taken from the real world or generated by an advanced
AI model?  Answer real if you think it is real, or fake if you
think it is AI-generated.  Respond with only one word, nothing
else.
```

Figure 7: Evaluation prompt used in general MLLMs.

Fake


```
{
    "messages": [
        {
            "role": "user",
            "content": "<image>\nIs this image real or generated by AI? If the image is real, output an empty JSON. If the image is tampered, locate the tampered object(s). For each tampered object, use its specific label (e.g., text content or general area describe) instead of a fix \"tampered\" label. Report the object name (label) and the bounding box coordinates (bbox_2d) in JSON format."
        }
    ],
    "solution": "fake",
    "images": [
        "/OpenMMSecV2/IMDL/non_generative/auto_pipeline/imgs/0abfd9bcf0c040cf8b3da5653dfb2d67.jpg"
    ],
    "meta_data": {
        "domain": "IMDL",
        "image": "0abfd9bcf0c040cf8b3da5653dfb2d67.jpg",
        "label": 1,
        "mani_type": "Non-Generative",
        "mask": "/OpenMMSecV2/IMDL/non_generative/auto_pipeline/masks/0abfd9bcf0c040cf8b3da5653dfb2d67.png",
        "ori_dataset": "tampercoco",
        "path": "/OpenMMSecV2/IMDL/non_generative/auto_pipeline/imgs/0abfd9bcf0c040cf8b3da5653dfb2d67.jpg",
        "real_source": "Coco",
        "sub_mani_type": "Auto-Pipeline",
        "bbox": [[0, 609, 9, 639], [93, 547, 114, 565], [0, 539, 15, 581], [91, 504, 192, 639], [0, 499, 28, 532], [73, 498, 89, 519], [302, 445, 377, 607], [326, 386, 332, 398], [426, 377, 479, 428], [333, 370, 418, 410], [128, 349, 135, 361], [215, 330, 239, 365], [197, 330, 216, 355], [254, 329, 321, 405], [152, 327, 165, 352], [104, 323, 122, 351], [55, 314, 96, 356]]
    }
}
```


Real


```
{
    "messages": [
        {
            "role": "user",
            "content": "<image>\nIs this image real or generated by AI? If the image is real, output an empty JSON. If the image is tampered, locate the tampered object(s). For each tampered object, use its specific label (e.g., text content or general area describe) instead of a fix \"tampered\" label. Report the object name (label) and the bounding box coordinates (bbox_2d) in JSON format."
        }
    ],
    "solution": "real",
    "images": [
        "/OpenMMSecV2/IMDL/generative/coco/imgs/f5d765dced8246cb8c622fbbb640e2dc.jpg"
    ],
    "meta_data": {
        "domain": "IMDL",
        "image": "f5d765dced8246cb8c622fbbb640e2dc.jpg",
        "label": 0,
        "mani_type": null,
        "mask": null,
        "ori_dataset": "Coco",
        "path": "/OpenMMSecV2/IMDL/generative/coco/imgs/f5d765dced8246cb8c622fbbb640e2dc.jpg",
        "real_source": "Coco",
        "sub_mani_type": null,
        "bbox": []
    }
}
```


Figure 6: The examples of openMMSec's training set.

```
You are a helpful assistant.  When the user asks a question, your
response must include two parts:  first, the reasoning process
enclosed in <think>...</think> tags, then the final answer enclosed
in <answer>...</answer> tags.  The critical answer or key result
should be placed within \\boxed{}.  Determine whether the image is
real or AI-generated/edited (fake), real or fake?
```

Figure 8: Evaluation prompt used in M2-Reasoning. The model lacks grounding ability AI et al. (2025), and to ensure the stability and accuracy of the classification results, the prompt does not inject the requirement for output bounding boxes.

```
Determine whether the image is real or AI-generated/edited; Outline
the position of each forgery area and output your answer directly.
If the image is AI-generated or edited, output:  <think>Your think
steps</think>fake, otherwise output <think> Your think steps
</think>real.  Please reasoning step by step.
```

Figure 9: Evaluation prompt used in Visionary-R1 and VL-Rethinker.

```
<image>Does the image looks real or fake?
```

Figure 10: Evaluation prompt used in FakeVLM.

```
You are an AI assistant that rigorously follows this response
protocol: 1. First, conduct a detailed analysis of the question.
Consider different angles, potential solutions, and reason through
the problem step-by-step. Enclose this entire thinking process
within <think> and </think> tags. 2. After the thinking section,
provide a clear, concise, and direct answer to the user's question.
Separate the answer from the think section with a newline. Ensure
that the thinking process is thorough but remains focused on the
query. The final answer should be standalone and not reference the
thinking section.
<image> Determine whether the image is real or AI-generated/edited
(fake); If the image is AI-generated or edited, answer is 'fake',
otherwise 'real'. Your answer must be 'fake' or 'real'
```

Figure 11: Evaluation prompt for authentic classification tasks used in InternVL3.5 Wang et al. (2025b).

```
<image> The image has been tampered with. Please provide the
bounding box coordinates of the region described in this sentence:
<ref>tampered area</ref>.
Please respond with a JSON object that contains the following
structure: { "bbox_2d": [x1, x2, y1, y2] } Ensure the 'bbox_2d'
value is a list of four integers representing the bounding box
coordinates.
```

Figure 12: Evaluation prompt used in InternVL3.5 Wang et al. (2025b). We refer to its training prompt for grounding tasks.

```
## System
You are an assistant trained to reason through image forensics.
When presented with an image and a question about whether it
is real or tampered, focus on reasoning through the process of
identifying potential manipulations.  Use the provided bounding
box coordinates to guide your decision-making without directly
referencing these coordinates.  Instead, refer to the region and
object described by the bounding box in general terms (e.g., 'the
area on the top left' or 'the central object').
Start by examining the image and look for signs of manipulation,
such as inconsistencies in texture, lighting, or object coherence.
Use your reasoning to explain why a certain area or object may be
tampered, describing it based on the bounding box's context.
## User Input
The image's tampered area:
'''json
{bboxes}
'''
```

Figure 13: The prompt of strong-to-weak distillation for tampered images.

```
## System
You are an assistant trained to reason through image forensics.
When presented with an image and a question about whether it
is real or tampered, focus on reasoning through the process of
identifying potential manipulations.  Use the provided label
('real' or 'AI-generated') to guide your decision-making.  use
general terms to describe regions and objects (e.g., 'the area on
the top left' or 'the central object').  Start by examining the
image and look for signs of manipulation, such as inconsistencies
in texture, lighting, or object coherence.  If the image is real,
provide reasoning based on the general characteristics of real
images, such as natural lighting, clear textures, and consistent
object proportions.  If the image is AI-generated, highlight
potential signs of tampering, such as unnatural textures, lighting
inconsistencies, or object incoherence in specific regions of
the image.  In either case, explain your reasoning and provide
a detailed description of the image's authenticity based on its
visual characteristics.

## User Input
The image is AI-generated.  Please examine the image for suspicious
areas that may have been tampered with.  Focus on regions with
unnatural textures, inconsistent lighting, or odd object shapes.
Based on your analysis, generate a bounding box (bbox_2d) for any
suspicious area you identify and provide a description of the
irregularity.  Your output should include the suspicious area
in the format [{'bbox_2d':  [x_min, y_min, x_max, y_max], 'label':
'suspicious area description'}].  Your output should include each
bounding box and its corresponding label in the format [{'bbox_2d':
[x_min, y_min, x_max, y_max], 'label':  'description'}, ...]
```

Figure 14: The prompt of strong-to-weak distillation for Ai-generated images.

```
## System
You are an assistant trained to reason through image forensics.
When presented with an image and a question about whether it
is real or tampered, focus on reasoning through the process of
identifying potential manipulations.  Use the provided label
('real' or 'AI-generated') to guide your decision-making.  use
general terms to describe regions and objects (e.g., 'the area on
the top left' or 'the central object').  Start by examining the
image and look for signs of manipulation, such as inconsistencies
in texture, lighting, or object coherence.  If the image is real,
provide reasoning based on the general characteristics of real
images, such as natural lighting, clear textures, and consistent
object proportions.  If the image is AI-generated, highlight
potential signs of tampering, such as unnatural textures, lighting
inconsistencies, or object incoherence in specific regions of
the image.  In either case, explain your reasoning and provide
a detailed description of the image's authenticity based on its
visual characteristics.
## User Input
The image is real.  Please explain why the image is authentic,
highlighting natural textures, lighting, and consistency in objects.
Point out any possible features that confirm it is real.
```

Figure 15: The prompt of strong-to-weak distillation for real images.

```
## System
You are an assistant trained to reason through image forensics.
When provided with an image and a list of bounding box coordinates,
your task is to describe each region corresponding to the bounding
box.  For each bounding box, you should reason through what object
or area the bounding box likely represents.  You should avoid
directly referencing the coordinates in the output.  Instead,
describe the object or area in general terms (e.g., 'the object
in the top-left corner' or 'the center object').  After reasoning,
return a list of dictionaries, each containing the bounding box
coordinates and a label for the object within that region (e.g.,
'person', 'vase', 'cup', or general descriptors like 'background
area').
## User Input
Given the image and the following bounding boxes:  ```json
{bboxes}
```, please identify and label the objects or areas corresponding to
each bounding box. The bounding box list is provided in the input,
and for each bounding box, provide a description or label based on
the content of the image. Your output should include each bounding
box and its corresponding label in the format [{'bbox_2d': [x_min,
y_min, x_max, y_max], 'label': 'description'}, ...].
```

Figure 16: The prompt of strong-to-weak distillation for region-level description.
```

Role: You are an expert AI Reasoning Auditor specializing in image forensics. Your task is to evaluate the logical consistency between a model's internal reasoning chain (Chain-of-Thought) and its final structured output.

Input Data:

- **User Query:** The forensic request (e.g., "Detect tampering in this image").
- **Ground Truth:** The actual label and bounding boxes of the image.
- **Model Response:** Contains two parts:
  1. <think>... </think>: The internal reasoning process.
  2. JSON Output: The final prediction (e.g., [{"label": "The description of the object", "bbox": [...] }, ...]).

Evaluation Criteria: You must classify the response into one of the following categories:

1. **Valid (Chosen):** The final JSON prediction matches the Ground Truth, AND the reasoning trace logically supports this conclusion (e.g., the model identifies artifacts in the think block and outputs the corresponding bbox).
2. **Inconsistent (Rejected - Hallucinated Reasoning):** The final JSON prediction is correct (matches Ground Truth), BUT the reasoning trace contradicts it (e.g., the think block states "the image looks authentic/natural," but the JSON outputs a tampering bbox to "hack" the reward).
3. **Incorrect (Rejected - Wrong Prediction):** The final prediction does not match the Ground Truth.

Output Format: Return a strictly valid JSON object:

```
{
    "status": "Valid" | "Inconsistent" | "Incorrect",
    "reasoning_summary": "Brief explanation of the judgment."
}
```

Figure 17: The instruction prompt used by the Judge Model to filter preference pairs for DPO training. We specifically target "Inconsistent" samples where the model achieves the correct answer via reasoning shortcuts or hallucinations.

E DATASET

We present data examples from the OpenMMSec training set (Tianchi, 2025). As illustrated in Figure 6, the dataset is categorized into fake and real classes, with original images sourced from multiple publicly available sub-datasets. Additionally, for entries containing mask fields, we extract bounding boxes (bboxes) to serve as the ground truth for model training and evaluation.

F PROMPT

The prompt used for training and inference of MMSec-R1is detailed in Figure 5. The evaluation prompts for other models are categorized as follows:

General MLLMs As shown in authentic classification Figures 7 and 11, grounding forgery Figure 12. we evaluate the general MLLMs prompt, including LLaVA-OneVision (An et al., 2025), InternVL3.5 (Wang et al., 2025b), and Qwen3-VL (Bai et al., 2025a).

Thinking MLLMs For authentic classification, as shown in Figures 8 and 9. we evaluate the thinking MLLMs prompt, including M2-Reasoning (AI et al., 2025), VL-Rethinker (Wang et al., 2025a), and Visionary-R1 (Xia et al., 2025).

Expert MLLMs For authentic classification, as shown in Figure 10. we evaluate the thinking MLLMs prompt, including FakeVLM (Wen et al., 2025).

Strong-to-weak Distillation To construct a high-quality cold-start dataset, we implement a comprehensive **Strong-to-weak Distillation Pipeline** leveraging Qwen-VL-Plus as the teacher model. We categorize the raw seed data into three distinct scenarios based on the nature of the forgery: (1) **Manipulated Images** (containing local forgery traces), (2) **AIGC Images** (fully synthetic content), and (3) **Real Images**. For **Manipulated Images**, ensuring the alignment between the textual reasoning and the visual forgery region is critical. As illustrated in Figure 13, we adopt a "guidance-based" generation strategy. Specifically, we inject the ground-truth bounding box coordinates into the prompt as explicit label signals. The teacher model is then instructed to focus its visual scrutiny on these specific coordinates and articulate the forensic evidence (e.g., splicing artifacts, noise inconsistency) found within that region. This ensures that the distilled Chain-of-Thought (CoT) data is spatially accurate and logically sound, effectively transferring the teacher's reasoning capability to the student model. In contrast, for **AIGC Images**, which typically exhibit holistic synthesis traces rather than local edits, we adopt a "discovery-based" strategy. As illustrated in Figure 14, we provide only the high-level class label "fake" without spatial hints. The teacher model is tasked with autonomously identifying characteristic AIGC artifacts (e.g., unnatural textures, semantic inconsistencies) and predicting the appropriate bounding boxes. This prevents the model from relying solely on provided coordinates and encourages the learning of intrinsic generative fingerprints. For **Real Images**, the goal is to enforce strict false-positive control. As shown in Figure 15, we provide the ground-truth label "real". The teacher model is instructed to verify the image's integrity and explicitly conclude that no manipulation exists, ensuring that the output strictly contains no bounding boxes. This establishes a baseline for authentic image verification and reduces model hallucination on benign content. Finally, to enhance the granularity of visual features, we employ the prompt detailed in Figure 16 to generate **Region-level Descriptions**. This step enriches the dataset by converting specific visual patterns within the identified bounding boxes into dense textual descriptions, further bridging the semantic gap between visual scrutiny and linguistic expression.

LIMITATIONS

Despite the promising results of MMSec-R1, several avenues for improvement remain: (1)**Model Scaling.** Our current work focuses on standard MLLMs (4B and 8B); extending the framework to larger Mixture-of-Experts (MoE) (Jiang et al., 2024) architectures is a key priority for future scaling; **Resolution Robustness.** We primarily use native resolutions. The robustness of the image encoder across diverse pixel densities and its impact on detection consistency require further investigation.