Cultivating Forensic Reasoning for Generalizable Multimodal Manipulation Detection

Yuchen Zhang[1] **Yaxiong Wang**[2*] **Kecheng Han**[1] **Yujiao Wu**[3]
Lianwei Wu[4] **Li Zhu**[1] **Zhedong Zheng**[5]
[1]School of Software Engineering, Xi'an Jiaotong University
[2]School of Computer Science and Information Engineering, Hefei University of Technology
[3]CSIRO [4]Northwestern Polytechnical University [5]University of Macau
yczhang@stu.xjtu.edu.cn wangyx@hfut.edu.cn

Abstract

Recent advances in generative AI have significantly enhanced the realism of multimodal media manipulation, thereby posing substantial challenges to manipulation detection. Existing manipulation detection and grounding approaches predominantly focus on manipulation type classification under result-oriented supervision, which not only lacks interpretability but also tends to overfit superficial artifacts. In this paper, we argue that generalizable detection requires incorporating explicit forensic reasoning, rather than merely classifying a limited set of manipulation types, which fails to generalize to unseen manipulation patterns. To this end, we propose **REFORM**, a reasoning-driven framework that shifts learning from outcome fitting to process modeling. REFORM adopts a three-stage curriculum that first induces forensic rationales, then aligns reasoning with final judgments, and finally refines logical consistency via reinforcement learning. To support this paradigm, we introduce **ROM**, a large-scale dataset with rich reasoning annotations. Extensive experiments show that REFORM establishes new state-of-the-art performance with superior generalization, achieving 81.52% ACC on ROM, 76.65% ACC on DGM4, and 74.9 F1 on MMFakeBench. The code is available at https://github.com/YcZhangSing/REFORM.

1 Introduction

The democratization of Generative AI (StabilityAI, 2023; Bai et al., 2023; Liu et al., 2023), has precipitated a paradigm shift in digital content creation. While this technological renaissance empowers creativity, it simultaneously lowers the barrier for fabricating hyper-realistic misinformation. The proliferation of sophisticated multimodal manipulations, ranging from subtle face swaps to fully synthesized news events, poses severe threats to information integrity and public trust (Pan et al., 2023; Park et al., 2022; Wang et al., 2025c). Consequently, detecting and grounding multimodal media manipulation (DGM4) has attracted wide attention and seen notable progress these years (Shao et al., 2024a; Zhang et al., 2025c; Liu et al., 2024).

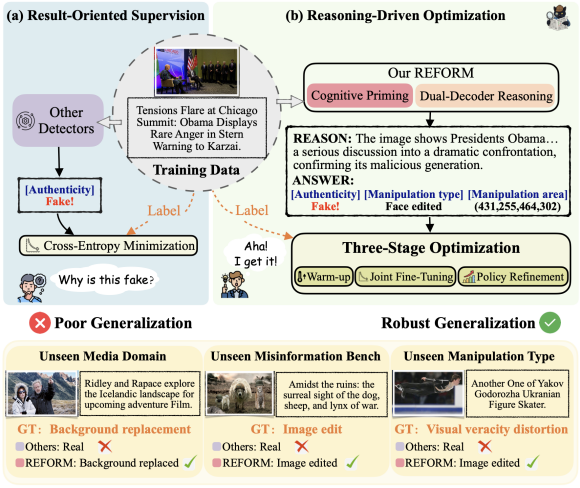


Figure 1: Comparison between learning paradigms. (a) The prevailing *Result-Oriented Supervision* usually suffers from poor generalization by merely fitting statistical artifacts of training data. (b) Our *Reasoning-Driven Optimization* facilitates robust generalization by explicitly optimizing the forensic reasoning chain, enabling the model to uncover intrinsic inconsistencies effectively across unseen domains.

Despite the progress made on the DGM4 problem, a pivotal challenge still remains: *generalization*. In practical scenarios, detectors must operate on diverse news domains and unseen manipulation patterns. Prior research has largely focused on invariant feature learning and knowledge augmentation (Liu et al., 2024; Zhang et al., 2025b,c). For example, FKA-Owl (Liu et al., 2024) augments Large Vision-Language Models with semantic correlations of forgery knowledge and visual artifacts for generalization enhancement. Concurrently, AMD (Zhang et al., 2025b) explicitly aligns visual artifacts with textual inconsistencies in a shared latent space to learn invariant forensic rep-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28545–28560

resentations robust to semantic contexts. While effective, these methods generally optimize models to map multimodal news directly to label annotations. This paradigm, which is heavily reliant on **result-oriented supervision**, tends to encourage models to fit specific statistical artifacts rather than cultivating a deep, transferable understanding of the underlying forensic logic.

As illustrated in Fig. 1 (a), relying solely on outcome supervision struggles to equip models with the intrinsic logic required for robust generalization. The essential reason is that optimizing exclusively for the final prediction (i.e., "Is this fake?") does not guarantee that the model implicitly learns the underlying forensic rationale, often leading to performance degradation when facing novel forgeries (Chu et al., 2025; Zhang et al., 2025a). To address this, we argue that the core of generalizable detection lies in explicitly cultivating forensic reasoning capabilities, as depicted in Fig.1 (b). Akin to a human forensic analyst, a model should not merely memorize the appearance of a forgery but rather deduce why an image-text pair is inconsistent through a logical chain of evidence. However, simply supervising this reasoning process is insufficient. While Supervised Fine-Tuning establishes a basic alignment, it remains constrained by the passive imitation of static annotations (Chu et al., 2025; Ma et al., 2025).

Motivated by this insight, we propose a new framework that shifts the learning objective from outcome fitting to **Reasoning-Driven Optimization**. To this end, we propose **R**easoning-**E**nhanced **F**orensic **O**ptimization via **R**einforcement **M**odeling (**REFORM**), a reasoning-endowed architecture designed to detect, ground, and explain multimodal manipulations. To effectively cultivate this capability, we implement a progressive three-stage learning curriculum. First, we perform *Cognitive Reasoning Warm-up*, teaching the model to articulate forensic rationales via data distillation. Second, we execute *Reasoning-Endowed Joint Fine-Tuning*, where the model learns to align its final judgment with its reasoning chain. Finally, recognizing that supervised learning suffers from exposure bias and lacks the ability to self-correct, we introduce *Constraint-Aware Policy Refinement*. Leveraging Reinforcement Learning with Group Relative Policy Optimization (Shao et al., 2024b), we incentivize the model to explore optimal reasoning paths that are logically consistent with the final verdict. This optimization strictly constrains the generation process with forensic accuracy, enabling REFORM to internalize robust judgment logic rather than fitting domain-specific patterns.

To rigorously benchmark this new paradigm, we construct **R**easoning-enhanced analysis for **O**mnibus **M**anipulation (**ROM**) dataset. Beyond the face-centric scope of previous benchmarks (Zhang et al., 2025b; Shao et al., 2023; Lian et al., 2024), ROM introduces scene-level synthesis and, crucially, provides detailed reasoning annotations for over **704k** samples to support process-oriented learning. In summary, our contributions are threefold:

- We identify the limitations of result-oriented manipulation detection and propose a paradigm shift towards reasoning-driven analysis, implementing REFORM, a new framework for robust multimodal manipulation detection, grounding, and forensic explanation.
- We introduce a progressive three-stage training framework culminating in a GRPO-based RL phase. This approach realizes Reasoning-Driven Optimization, effectively aligning the model's cognitive process with forensic logic to distinguish intrinsic anomalies from statistical biases.
- We curate ROM, a large-scale, comprehensive benchmark that includes omnibus manipulation beyond face-related manipulations and provides high-quality reasoning supervision, setting a new standard for interpretable multimodal forensics.

2 Related Work

Multimodal Misinformation Detection. Given challenging aligned forgeries (Shao et al., 2024a), recent approaches typically leverage external knowledge or refine internal features. For instance, FKA-Owl (Liu et al., 2024) utilizes LVLMs for factual verification, while RamDG (Shen et al., 2025) cross-references news with attribute databases. Conversely, feature-centric methods, *e.g.*, AMD (Zhang et al., 2025b), encode artifact tokens alongside semantic content. However, these methods predominantly rely on *result-oriented supervision*. This paradigm usually encourages shortcut learning (Zhang et al., 2025a), causing models to overfit statistical artifacts rather than logical evidence, which degrades generalization on unseen domains (Chu et al., 2025).

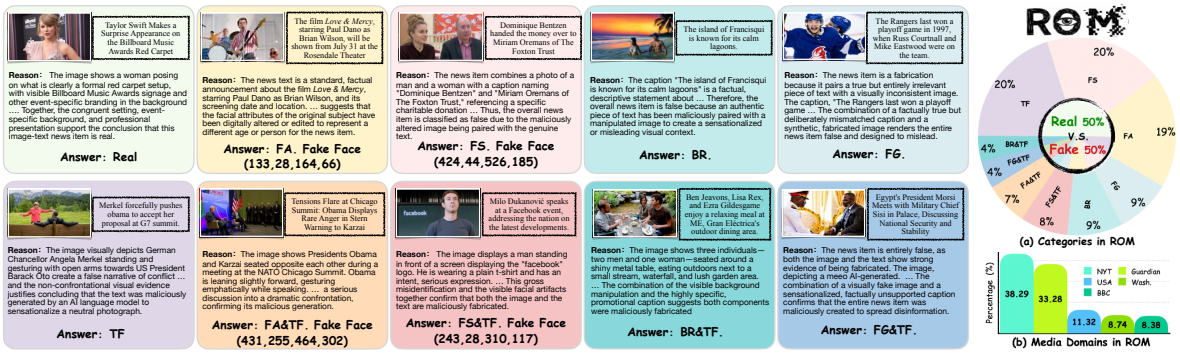


Figure 2: Overview of the ROM dataset. *Left:* Representative samples spanning 9 manipulated and 1 real categories, ranging from face-centric edits to scene-level synthesis, each accompanied by a detailed reasoning annotation. *Right:* Statistical distribution showing the diversity of manipulation types and the coverage of news media domains.

Vision Language Models and Reasoning Optimization. Large Vision Language Models, such as Qwen-VL (Bai et al., 2025) and InternVL (Wang et al., 2025d), have revolutionized multimodal understanding and been adapted for forensics via Supervised Fine-Tuning (SFT) (Yan et al., 2025; Wang et al., 2025b,a). However, SFT often fails to cultivate robust reasoning as it focuses on mimicking outputs, limiting generalization and self-correction (Huang et al., 2025a; Chu et al., 2025). To address this, *Reasoning-Driven Optimization* utilizing Chain-of-Thought (Li et al., 2025; Zhang et al., 2025a) and Reinforcement Learning methods like GRPO (Shao et al., 2024b) has emerged to incentivize the reasoning process itself (Huang et al., 2025b; Zheng et al., 2025). Inspired by this, our REFORM framework integrates RL-based optimization to enforce logical consistency between forensic reasoning chains and verdicts, enabling robust and generalizable detection.

3 Methodology

3.1 Data Preparation

To support our process reasoning-focused training as well as construct a comprehensive benchmark for generalization evaluation, we prepare Reasoning-enhanced analysis for Omnibus Manipulation (ROM) dataset, a full-scene large-scale benchmark with thinking annotation. While incorporating face-centric samples from MDSM (Zhang et al., 2025b), ROM significantly transcends previous boundaries by expanding the scope to scene-level synthesis and integrating logical reasoning. As illustrated in Fig. 2, the ROM dataset comprises **704,456** image-text pairs across five news domains and nine manipulation categories.

Manipulation Scope Expansion. ROM assimilates six face-specific classes from MDSM (Zhang et al., 2025b), comprising *Original* (Orig), *FaceAttribute* (FA), *FaceSwap* (FS), *TextFabrication* (TF), and their composites (FA&TF, FS&TF). Crucially, we extend this taxonomy by introducing four scene-level categories: *BackgroundReplacement* (BR), *FullGeneration* (FG), and their text-fabricated variants (BR&TF, FG&TF). This expansion moves beyond local facial edits to holistic scene synthesis, utilizing diverse generative models (e.g., SD3 (Esser et al., 2024), FLUX.1 (Labs et al., 2025)) to ensure artifact diversity.

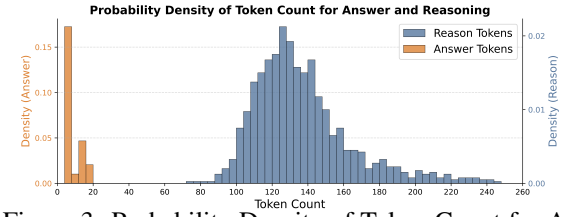


Figure 3: Probability Density of Token Count for Answer and Reasoning.

Reasoning Data Distillation. We augment ROM with rationale annotations to enhance interpretability. Using InternVL3.5-30B (Wang et al., 2025d), we generate textual reasoning for each image-text pair given its manipulation label. Fig. 2 visualizes these reasoning samples. As shown in Fig. 3, generated rationales typically peak around 130 tokens, offering significantly richer context than standard short answers (<20 tokens).

3.2 Architecture

Fig. 4a illustrates the architecture of our proposed REFORM model, which ingests multimodal inputs and produces both detection and grounding results in textual form. REFORM is a sequence-to-sequence framework employs a novel encoder-dual-decoder structure as its backbone. We design a *Cognitive Priming Module* to enhance the model's ability to capture forgery-related cues, and a *Dual-Decoder* to strengthen its capability in interpreting forged evidence.

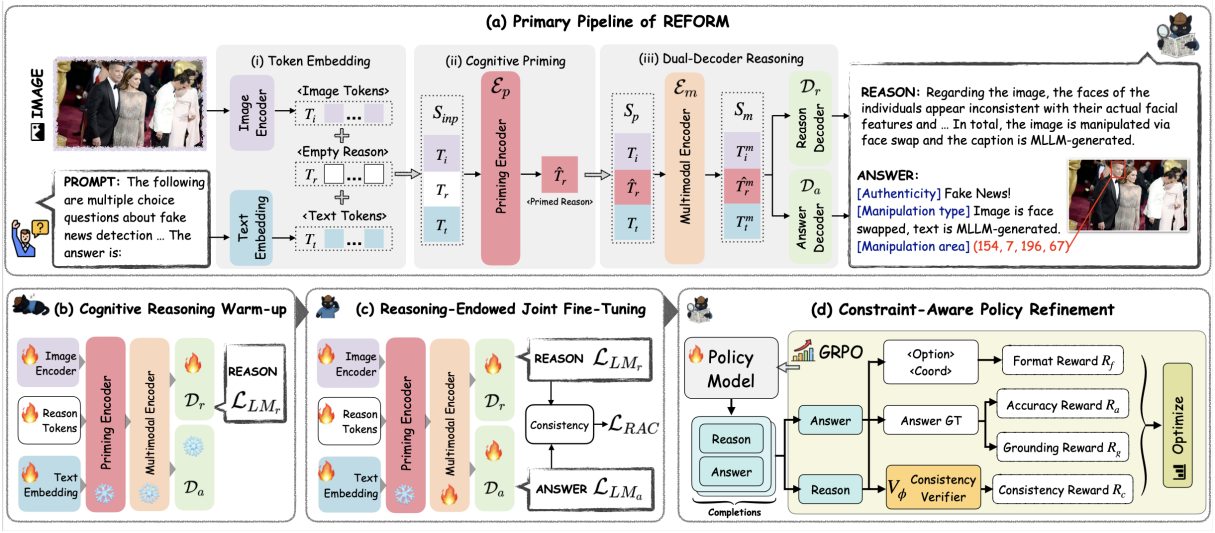


Figure 4: Overview of the REFORM framework and its three-stage training curriculum. (a) The primary pipeline employs a Cognitive Priming Encoder $\mathcal{E}_p$ and a Dual-Decoder structure, $\mathcal{D}_r$ and $\mathcal{D}_a$, for reasoning-driven detection. (b) Cognitive Reasoning Warm-up via partial freezing. (c) Reasoning-Endowed Joint Fine-Tuning incorporating the Reason-Answer Consistency Loss $\mathcal{L}_{RAC}$. (d) Constraint-Aware Policy Refinement using GRPO-based Reinforcement Learning to align forensic logic with the final verdict.

Prompt Paradigm. The Prompt follows heuristic question(human)-answer(assistant) paradigm where the image-question pair serves as input and the text response constitutes REFORM's output

Prompt Template Definition

###Human: ⟨*Task*⟩⟨*Options*⟩⟨*Grounding*⟩
###Assistant: ⟨*Response*⟩[⟨*Coordinates*⟩]

- ⟨*Task*⟩: Specifies manipulation detection objective & pairs input.
- ⟨*Options*⟩: Lists all candidate answers for the ROM task.
- ⟨*Grounding*⟩: Triggers region localization conditionally.
- ⟨*Response*⟩: Encapsulates the correct answers.
- [⟨*Coordinates*⟩]: Encloses edited region coordinates.

Token Embedding. Through the image encoder, the input image is transformed into a sequence of <*Image Tokens*>, denoted as T_i, where n_v is the number of visual tokens and d is the embedding dimension. Similarly, the question prompt is passed through an embedding layer to produce a sequence of <*Text Tokens*>, represented as T_t.

Cognitive Priming. This module is designed to perceive manipulation-related cues from the input data and distill these forensic signals into the Cognitive token sequence T_r. Structurally, the module is built upon the Cognitive Priming Encoder $\mathcal{E}_p$, instantiated as a standard Transformer encoder. Crucially, $\mathcal{E}_p$ operates in a parameter-frozen state; this constraint forces the learnable query tokens T_r to actively extract and aggregate multimodal inconsistencies from the fixed semantic context provided by T_i and T_t. Formally, the tokens are concatenated to form $S_{inp} = [T_i; T_r; T_t]$, where $[\cdot;\cdot]$ denotes concatenation along the token dimension. The sequence S_{inp} is then processed by $\mathcal{E}_p$:

$$\mathcal{E}_p(S_{inp}) = [\hat{T}_i; \hat{T}_r; \hat{T}_t]. \quad (1)$$

We retain only $\hat{T}_r$ for further processing, which encapsulates the distilled manipulation signals.

Dual-Decoder Reasoning. The primed reason tokens $\hat{T}_r$ are concatenated with the initial image tokens T_i and text tokens T_t to construct the composite input sequence $S_p = [T_i; \hat{T}_r; T_t]$. This sequence is encoded by the multimodal encoder $\mathcal{E}_m$ to yield the latent representation S_m. Subsequently, S_m is fed into the Reason-Answer Dual-Decoder in parallel to generate textual responses. The Answer Decoder $\mathcal{D}_a$ generates a structured textual output comprising three distinct predictions, as shown in Fig. 4a: (1) the authenticity classification (e.g., identifying the news as 'fake'), (2) fine-grained manipulation types (e.g., detecting face-swapped and AI-generated captions), and (3) manipulation localization (e.g., outputting coordinates to pinpoint the manipulated region). Meanwhile, the Reason Decoder $\mathcal{D}_r$ is tasked with producing a comprehensive explanation that rationalizes the model's verdict.

We adopt a dual-decoder architecture rather than a shared one for two reasons: (1) The expanded parameter space offers greater flexibility for optimizing both tasks independently, thereby reducing

potential gradient conflicts during joint training. (2) It supports seamless switching between two modes: plugging reasoning decoder for reasoning mode, while un-equipping it for answer-only mode, resulting in a more configurable framework.

3.3 Cognitive Reasoning Warm-up

This initial phase is dedicated to aligning the model's cognitive process with the ground-truth forensic reasoning. Our primary objective is to specifically enhance the forensic reasoning capability by supervising the model with the distilled rationale annotations.

As illustrated in Fig. 4b, we adopt a partial freezing strategy where the Multimodal Encoder $\mathcal{E}_m$, the Answer Decoder $\mathcal{D}_a$, and the Cognitive Priming Encoder $\mathcal{E}_p$ remain frozen. The optimization exclusively targets the learnable reason tokens T_r and the Reason Decoder $\mathcal{D}_r$. Under the supervision of the ground-truth rationales, the model is trained to reconstruct the detailed forensic explanation. And the reason token T_r are forced to extract and encode specific manipulation traces that match the logical patterns in the reasoning annotations. We employ the standard causal language modeling objective as the reasoning generation loss, denoted as $\mathcal{L}_{LM_r}$, to guide this alignment:

$$\mathcal{L}_{LM_r} = -\frac{1}{L}\sum_{t=1}^{L}\log P(y_t \mid y_{<t}, S_m), \quad (2)$$

where L is the length of the ground-truth rationale sequence, and y_t denotes the t-th target token.

3.4 Reasoning-Endowed Joint Fine-Tuning

Having established a coherent forensic logic foundation, we transition to the Supervised Fine-Tuning stage to endow the model's judgment logic with these forensic reasoning capabilities. This phase is executed through two key strategies:

Dual-Stream Generative Optimization. First, we activate the model's full capacity for simultaneous reasoning and judgment. As shown in Fig. 4c, we unfreeze the entire backbone, including $\mathcal{E}_m$ and D_a. The model is now tasked with generating both the reasoning chain R and the structured answer A. Thus, we optimize both streams via language modeling losses $\mathcal{L}_{LM_r}$ and $\mathcal{L}_{LM_a}$, where $\mathcal{L}_{LM_a}$ mirrors Eq. 2 on answers annotations.

Reason-Answer Semantic Alignment. Solely minimizing generative losses independent of each other could lead to logical discrepancies, where the generated reasoning contradicts the final verdict. To bridge this potential semantic gap, we introduce the *Reason-Answer Consistency Loss* ($\mathcal{L}_{RAC}$). This objective enforces a minimum semantic similarity between the representation of the rationale and the answer, ensuring the reasoning trajectory effectively substantiates the final answer.

We derive fixed-size global embeddings $\mathbf{v}^R$ and $\mathbf{v}^A$ via mean pooling over hidden states, masked to exclude padding tokens to capture only valid semantic content. We then employ a margin-based hinge loss to penalize alignments falling below a threshold η:

$$\mathcal{L}_{RAC} = \max\{0, \eta - \cos(\mathbf{v}^R, \mathbf{v}^A)\}. \quad (3)$$

Overall, the full objectives for this Reasoning-Endowed Joint Fine-Tuning stage is:

$$\mathcal{L}_{RJF} = \mathcal{L}_{LM_r} + \mathcal{L}_{LM_a} + \mathcal{L}_{RAC}. \quad (4)$$

3.5 Constraint-Aware Policy Refinement

To mitigate SFT's exposure bias and encourage the exploration of optimal reasoning, we adopt Group Relative Policy Optimization (GRPO) with a multi-dimensional reward function $\mathcal{R}$ consisting of four components:

Consistency Reward ($\mathcal{R}_c$) measures the semantic alignment between the generated reasoning chain R and the predicted manipulation types, denoted as $\hat{c}_i$ (image-modal) and $\hat{c}_t$ (text-modal). We introduce a *Consistency Verifier* V_ϕ to evaluate the logical entailment between the reasoning and these answer components. We instantiate V_ϕ using a lightweight TinyBERT encoder (Jiao et al., 2020) equipped with two parallel classification heads for both modalities. Prior to RL training, V_ϕ is pre-trained on ground-truth reason-label pairs, achieving over 99% classification accuracy. This ensures V_ϕ can reliably deduce the correct manipulation category solely from a reasoning description. The consistency reward is calculated by checking if V_ϕ's deductions match the predictions from $\mathcal{D}_a$:

$$\mathcal{R}_c = \mathbb{I}(V_\phi^i = \hat{c}_i) + \mathbb{I}(V_\phi^t = \hat{c}_t), \quad (5)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Accuracy Reward ($\mathcal{R}_a$) aligns predictions with ground-truth labels via binary verification ($\mathcal{R}_{bin}$) and fine-grained type recognition ($\mathcal{R}_{fin}$). Given ground truths y_{bin}, c_i, c_t and predictions $\hat{y}, \hat{c}_i, \hat{c}_t$,

we formulate the reward as:

$$\mathcal{R}_{bin} = \mathbb{I}(\hat{y} = y_{bin}), \tag{6}$$

$$\mathcal{R}_{fin} = \mathbb{I}(\hat{c}_i = c_i) + \mathbb{I}(\hat{c}_t = c_t), \tag{7}$$

$$\mathcal{R}_a = \mathcal{R}_{bin} + \mathcal{R}_{fin}. \tag{8}$$

Grounding Reward ($\mathcal{R}_g$) evaluates the spatial precision by calculating the Intersection over Union (IoU) between the predicted bounding boxes $\hat{b}$ and the ground truth boxes b:

$$\mathcal{R}_g = \text{IoU}(\hat{b}, b). \tag{9}$$

Format Reward ($\mathcal{R}_f$) enforces strict adherence to the specified output structure, as defined in Sec. 3.2. Let $\hat{a}$ be the generated answer string and $\mathcal{S}$ be the set of valid regex patterns:

$$\mathcal{R}_f = \mathbb{I}(\hat{a} \in \mathcal{S}). \tag{10}$$

Optimization Objective. For each input prompt x, we sample a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ from the current policy π_θ. The total reward for the i-th output is the sum of the above components: $\mathcal{R}^{(i)}_{total} = \mathcal{R}_c + \mathcal{R}_a + \mathcal{R}_g + \mathcal{R}_f$. To compute the advantage A_i, we normalize the rewards within the group to reduce variance. Let $r_i = \frac{\pi_\theta(o_i|x)}{\pi_{\text{old}}(o_i|x)}$ denote the probability ratio. The surrogate objective for the i-th sample is formulated as:

$$\mathcal{J}^{(i)}_{\text{clip}} = \min\left(r_i A_i, \text{clip}\left(r_i, 1 \pm \epsilon\right) A_i\right). \tag{11}$$

The final GRPO objective averages this surrogate objective while applying a KL-divergence penalty to prevent policy collapse:

$$\mathcal{L}_{\text{G}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \mathcal{J}^{(i)}_{\text{clip}} - \beta \mathbb{D}_{KL}(\pi_\theta || \pi_{\text{ref}})\right]. \tag{12}$$

4 Experiment

Please refer to Appendix for implementation details and evaluation metric. Since rationale prediction is introduced solely to cultivate reasoning capabilities without ground-truth rationales, we validate its effectiveness implicitly through the performance of the primary detection and grounding tasks.

4.1 Quantitative Results

We evaluate REFORM on three comprehensive benchmarks: ROM, MMFakeBench (Liu et al., 2025), and DGM4 (Shao et al., 2024a). **(1) Cross-Domain Generalization (Tab. 1):** On ROM, REFORM significantly outperforms state-of-the-art (SOTA) baselines. In the challenging cross-domain setting (Train on NYT), REFORM achieves an average accuracy of 88.22, surpassing AMD (85.92) and HAMMER (72.41). Notably, REFORM also outperforms MMD-Agent-34B (57.45), an agentic pipeline that utilizes iterative reasoning and external knowledge (Wikipedia) retrieval. **(2) Zero-shot Generalization (Tab. 2):** Despite dealing with unseen manipulation types (e.g., manual PS editing) in MMFakeBench, REFORM achieves a remarkable F1 score of 74.9. It significantly outperforms 7B and 13B parameter LVLMs, demonstrating that forensic reasoning equips small-scale models (0.3B) with superior zero-shot generalizability compared to their larger counterparts. **(3) Superiority over Specialized Detectors (Tab. 3):** REFORM establishes a new SOTA on the face-centric DGM4 benchmark. As observed, fine-tuned LVLMs struggle to detect subtle artifacts, yielding unsatisfactory mAP (< 47). While specialized detectors like AMD and FKA-Owl, REFORM significantly outperforms them with average mAP 65.72. It shows that optimizing forensic reasoning captures intrinsic manipulation traces with greater cross-domain universality compared to feature-engineering approaches, leading to superior generalization.

4.2 Ablation Studies

We conduct extensive ablation studies, as summarized in Tab. 4. Note that all data presented in this table denote the cross-domain average performance.

Impact of Component Modules. Tab. 4a validates our three-stage curriculum. The base model (LM_a) achieves 84.88 accuracy on NYT. Introducing the reasoning objective (LM_r) improves performance to 87.76, confirming that rationalization aids detection. The addition of consistency loss (RAC) and GRPO further boosts accuracy to 88.22 and mIoU to 78.48, demonstrating that enforcing logical consistency between reasoning and judgment is crucial for optimal performance.

Impact of Reason Token Length. We examine the effect of reason token length on cross-domain average performance. in Tab. 4b. Performance initially improves with length as the model captures more forensic details, peaking at 32 tokens with optimal ACC 88.22.

Sensitivity Analysis of Consistency Margin. As shown in Tab. 4c on η in $\mathcal{L}_{RAC}$. We observe that the model exhibits robustness within a reasonable range, and $\eta = 0.0$ yields the best trade-off, effectively penalizing semantic misalignment without

Table 1: Comparison of multimodal learning methods on ROM, where the background gray indicates the intra-domain test. The better results are in **bold**. AVG refers to the average performance across 5 news domains.

Setting	Method	Test Domain																	
		NYT			Guardian			USA			Wash.			BBC			AVG		
		ACC	mAP	mIoU	ACC	mAP	mIoU	ACC	mAP	mIoU	ACC	mAP	mIoU	ACC	mAP	mIoU	ACC	mAP	mIoU
Zero-Shot	Yi-VL-6B (01.AI, 2024)	31.03	21.86	2.67	20.13	19.18	1.14	19.65	12.76	1.72	26.15	25.47	1.79	29.67	17.38	1.11	25.33	19.33	1.69
	DeepSeek-VL2-27B (Wu et al., 2024)	45.74	34.01	8.21	39.76	27.67	7.13	31.78	36.91	7.21	29.25	29.03	6.90	33.01	29.62	4.03	35.91	31.45	6.70
	LLaVA-v1.6-34B (Liu et al., 2023)	47.66	41.09	7.45	38.54	31.12	6.76	32.89	37.27	6.43	31.09	31.23	6.78	36.00	29.58	4.12	37.24	34.06	6.31
	Qwen2.5-VL-72B (Bai et al., 2025)	50.74	42.24	12.79	40.18	32.70	11.08	33.64	37.60	10.99	35.11	34.29	10.28	37.11	30.51	7.69	39.35	35.47	10.56
	GPT-4o (Hurst et al., 2024)	**51.98**	42.15	**16.42**	**42.18**	32.09	**11.79**	46.08	38.31	14.36	39.18	33.41	11.07	37.21	30.72	10.47	43.33	35.34	12.82
	Gemini-2.5 (Comanici et al., 2025)	50.86	**42.73**	15.41	42.11	**33.16**	11.54	**49.42**	**40.56**	**15.24**	**39.84**	**38.70**	**11.87**	**39.54**	**34.84**	**11.71**	**44.35**	**38.00**	**13.15**
	MMD-Agent-34B (Liu et al., 2025)	61.34	41.63	40.46	61.13	36.24	55.67	62.32	47.11	63.54	51.23	42.42	63.07	51.23	45.34	71.47	57.45	42.55	58.84
Train on NYT	Fine-tuned LVLMs																		
	Qwen2.5-3B (Bai et al., 2025)	86.76	57.54	65.34	80.04	41.34	62.78	72.07	34.82	62.59	74.87	38.43	55.75	73.13	39.98	70.00	77.38	42.42	63.29
	LLaVa-v1.6-7B (Liu et al., 2023)	94.42	81.23	83.87	83.45	60.37	58.06	70.16	53.96	61.36	69.80	53.13	59.27	84.59	52.26	73.99	80.48	60.19	67.31
	Deepfake Detectors																		
	ViLT (Kim et al., 2021)	80.17	69.71	32.07	67.02	55.16	30.61	66.14	56.71	31.18	71.71	58.31	29.61	71.03	45.11	30.62	71.21	57.00	30.82
	HAMMER (Shao et al., 2023)	81.49	73.97	59.53	67.94	56.24	43.29	68.05	61.35	49.12	70.94	57.61	49.02	73.63	54.50	45.05	72.41	60.73	49.20
	HAMMER++ (Shao et al., 2024a)	82.91	74.07	59.92	66.42	57.04	44.36	67.55	63.22	51.04	71.11	58.12	49.62	74.98	52.01	45.39	72.59	60.89	50.07
	FKA-Owl (Liu et al., 2024)	95.76	89.66	74.13	82.22	64.42	65.17	**83.64**	63.33	73.74	78.37	59.84	64.52	85.83	61.03	69.13	85.17	67.66	69.34
	AMD (Zhang et al., 2025b)	94.13	89.28	**89.22**	85.52	68.98	75.17	81.32	67.72	73.74	80.14	65.01	74.15	88.48	64.68	74.64	85.92	71.13	77.38
	REFORM (Ours)	**96.69**	**91.76**	88.34	**86.87**	**73.18**	**76.86**	83.35	**69.16**	**75.26**	**83.87**	**72.15**	**76.40**	**90.34**	**74.17**	**75.52**	**88.22**	**76.08**	**78.48**
Train on Guardian	Fine-tuned LVLMs																		
	Qwen2.5-3B (Bai et al., 2025)	61.90	22.23	61.99	92.76	75.62	77.23	67.59	44.62	63.44	70.76	36.04	67.64	76.21	49.17	72.38	73.84	45.54	68.54
	LLaVa-v1.6-7B (Liu et al., 2023)	63.15	23.14	65.35	93.66	75.13	83.37	67.68	45.87	66.72	69.08	33.74	62.54	80.63	54.56	77.52	74.84	46.49	71.10
	Deepfake Detectors																		
	ViLT (Kim et al., 2021)	68.06	47.36	38.44	88.50	89.79	60.36	65.37	51.17	46.13	63.30	58.34	47.02	78.89	49.15	58.69	72.82	59.16	50.13
	HAMMER (Shao et al., 2023)	71.15	49.74	47.24	90.69	92.50	69.25	63.50	51.47	57.41	64.12	57.02	58.42	80.04	43.43	73.26	73.90	58.83	61.12
	HAMMER++ (Shao et al., 2024a)	70.79	47.57	47.61	89.63	91.90	71.67	61.53	50.85	58.39	62.58	55.17	59.66	81.14	43.91	73.35	73.13	57.88	62.14
	FKA-Owl (Liu et al., 2024)	**75.46**	45.14	56.41	91.83	70.28	72.51	70.32	30.65	76.85	70.06	31.56	72.34	83.14	42.93	63.16	78.16	44.11	68.25
	AMD (Zhang et al., 2025b)	74.23	42.98	64.49	91.07	76.54	90.16	71.50	35.76	80.42	71.46	37.84	80.94	**83.42**	46.18	76.13	78.34	47.86	78.43
	REFORM (Ours)	74.38	**63.62**	**67.53**	**94.04**	**92.90**	**94.07**	**79.16**	**58.48**	**84.56**	**78.04**	**59.49**	**83.41**	81.98	**64.26**	**77.75**	**81.52**	**67.75**	**81.46**

Table 2: Comparison of zero-shot binary detection performance on the MMFakeBench validation and test sets. We compare our proposed REFORM with baseline models using "Standard" and "MMD-Agent" prompting paradigms as defined in the MMFakeBench paper. Baseline results are cited from the original MMFakeBench publication.

Model	Language	Prompt	Validation (1000)				Test (10000)			
Name	Model	Method	F1	Precision	Recall	ACC	F1	Precision	Recall	ACC
LVLMs with 7B Parameter										
InstructBLIP (Dai et al., 2023a)	Vicuna-7B (Dai et al., 2023b)	Standard	14.7	30.8	13.2	8.1	16.1	40.5	14.2	8.8
Qwen-VL (Bai et al., 2023)	Qwen-7B (Bai et al., 2023)	Standard	43.6	50.6	44.9	60.3	44.0	51.6	45.2	60.5
PandaGPT (Su et al., 2023)	Vicuna-7B (Dai et al., 2023b)	Standard	24.6	60.6	50.5	30.9	24.1	61.7	50.4	30.6
mPLUG-Owl2 (Ye et al., 2025)	LLaMA2-7B (Touvron et al., 2023)	Standard	47.2	64.9	52.3	70.6	48.7	71.1	53.3	**71.4**
LLaVA-1.6 (Liu et al., 2023)	Vicuna-7B (Dai et al., 2023b)	Standard	48.1	48.2	48.5	59.5	52.5	53.0	52.6	62.5
LVLMs with 13B Parameter										
InstructBLIP (Dai et al., 2023a)	Vicuna-13B (Dai et al., 2023b)	Standard	41.1	35.0	49.9	69.9	41.1	35.0	49.9	69.8
		MMD-Agent	51.3	53.4	54.0	53.1	47.9	50.1	50.1	49.9
LLaVA-1.6 (Liu et al., 2023)	Vicuna-13B (Dai et al., 2023b)	Standard	41.1	35.0	50.0	69.7	42.3	57.3	50.1	69.5
		MMD-Agent	51.8	66.7	54.6	**71.4**	50.2	67.3	53.9	71.3
Ours										
REFORM(Ours)	Florence2-0.3B (Bin et al., 2023)	Ours (Sec.3.2)	**74.9**	**74.5**	**75.4**	64.7	**74.1**	**74.1**	**74.0**	63.7

disrupting the learning of distinct modal features.

Impact of GRPO Reward Configurations. As illustrated in Tab. 4d, we analyze the cumulative effect of reward components. I denotes the baseline model without RL. Subsequent configurations sequentially introduce the Format Reward ($\mathcal{R}_f$, II), Accuracy Reward ($\mathcal{R}_a$, III), Grounding Reward ($\mathcal{R}_g$, IV), and Consistency Reward ($\mathcal{R}_c$, V). Observing the Guard. ACC, NYT mAP, and Guard. mAP metrics, a substantial performance leap occurs at III with the introduction of $\mathcal{R}_a$. Similarly, the integration of $\mathcal{R}_c$ in V yields another distinct performance boost. Regarding NYT ACC, since the baseline performance is already saturated (87.84), the scope for RL-driven improvement is naturally limited; thus, $\mathcal{R}_a$ and $\mathcal{R}_c$ yield only marginal gains, which is expected. Meanwhile, the inclusion of $\mathcal{R}_g$ significantly enhances mIoU across both domains.

Impact of RL Training Configurations. We analyze the impact of generation group size G throughout the training process in Tab. 4e-f, with performance measured by mAP. All RL configurations significantly outperform the SFT baselines (73.25

Table 3: Comparison of multimodal learning methods on DGM4 (%), where the guardian domain with background gray is intra-domain. P_{tok} is Precision of fake token grounding.

Method	Test Domain																			
	Guardian				USA				Wash.				BBC				AVG			
	ACC	mAP	P_{tok}	mIoU	ACC	mAP	P_{tok}	mIoU	ACC	mAP	P_{tok}	mIoU	ACC	mAP	P_{tok}	mIoU	ACC	mAP	P_{tok}	mIoU
Fine-tuned LVLMs																				
Qwen2.5-3B (Bai et al., 2025)	61.57	37.36	61.35	40.19	60.65	34.25	70.19	33.28	51.79	32.23	63.23	35.11	62.20	38.02	61.22	41.23	59.05	35.47	64.00	37.45
LLaVa-v1.6-7B (Liu et al., 2023)	68.67	46.26	65.71	42.30	62.54	37.48	71.24	35.63	63.16	40.27	71.03	34.22	66.14	46.44	62.17	42.18	65.13	42.61	67.54	38.58
Deepfake Detectors																				
ViLT (Kim et al., 2021)	68.27	42.29	69.87	43.19	52.79	31.28	62.11	33.78	55.76	33.26	57.17	31.10	44.14	39.68	59.06	21.96	55.24	36.63	62.05	32.51
HAMMER (Shao et al., 2023)	78.34	66.79	78.27	61.09	64.97	40.49	73.76	40.51	63.54	40.26	76.13	38.53	54.97	40.84	81.48	43.74	65.46	47.10	77.41	45.97
HAMMER++ (Shao et al., 2024a)	79.13	67.11	78.24	62.15	65.25	40.74	73.24	41.14	63.83	40.34	76.17	38.21	54.24	41.25	81.73	43.23	65.61	47.36	77.35	46.18
FKA-Owl (Liu et al., 2024)	82.97	53.86	87.70	65.69	67.57	38.97	79.44	32.57	67.05	37.70	81.55	31.86	70.26	40.20	84.54	46.48	71.96	42.68	83.31	44.15
AMD (Zhang et al., 2025b)	84.61	68.50	82.78	81.24	70.62	43.20	75.73	41.99	70.28	43.36	77.76	39.05	**72.37**	56.57	83.76	45.20	74.47	52.91	80.01	51.87
REFORM (ours)	**91.10**	**84.30**	**89.92**	**83.88**	**71.95**	**59.08**	**81.23**	**55.19**	**72.84**	**59.53**	**83.13**	**53.79**	70.70	**59.97**	**86.37**	**47.92**	**76.65**	**65.72**	**85.17**	**60.19**

Table 4: Ablation of components (a), reason token length (b), η sensitivity (c), reward (d), and completion (e-f).

Components				NYT			Guardian		
LM_a	LM_r	RAC	GRPO	ACC	mAP	mIoU	ACC	mAP	mIoU
✓				84.88	66.16	75.98	72.18	45.86	78.72
✓	✓			87.76	73.01	77.68	74.74	53.65	79.59
✓	✓	✓		87.84	73.25	78.00	75.71	54.11	79.58
✓	✓	✓	✓	**88.22**	**76.08**	**78.48**	**81.52**	**67.75**	**81.64**

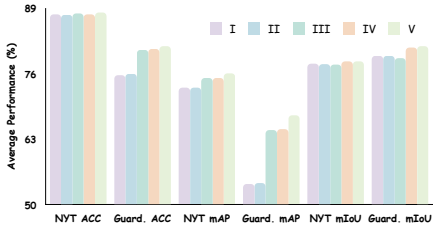
(a) Components Ablation.

Len.	NYT			Guardian		
	ACC	mAP	mIoU	ACC	mAP	mIoU
16	87.49	75.10	76.83	80.78	67.26	80.38
32	**88.22**	76.08	**78.48**	**81.52**	**67.75**	**81.46**
64	88.16	**76.17**	78.28	81.39	67.52	81.41

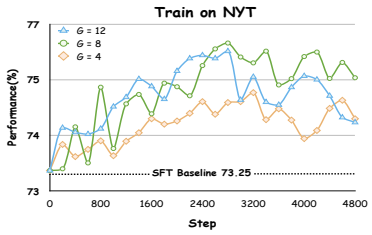
(b) Reason Token Length.

η	NYT			Guardian		
	ACC	mAP	mIoU	ACC	mAP	mIoU
-0.1	88.12	75.50	78.03	81.11	67.11	81.18
0.0	**88.22**	**76.08**	**78.48**	81.52	**67.75**	**81.46**
0.1	88.05	75.87	78.08	**81.94**	67.67	81.19

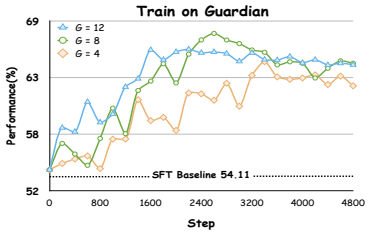
(c) Sensitivity of η.



(d) GRPO Reward Configurations.



(e) GRPO training on NYT.



(f) GRPO training on Guardian.

Table 5: Efficiency comparison. We report total parameters and throughput (pairs/sec) on RTX 4090. Fast Mode refers to REFORM generating only prediction labels without the reasoning chain.

Method	Params (M) ↓		Throughput (p/s) ↑	
	Total	Trainable	Train	Inference
ViLT	121.07	121.07	1.85	2.38
HAMMER(++)	441.12	228.25	28.97	61.28
FKA-Owl	6771.98	33.55	1.25	1.33
MMD-Agent	34751.17	34447.66	-	0.02
AMD	276.95	276.95	5.55	13.38
REFORM (Ours)	**376.23**	**376.23**	**4.68**	*Explainable*: 1.03 ***Fast Mode*: 13.17**

on NYT and 54.11 on Guardian), strongly validating our policy refinement. Notably, $G = 8$ delivers the most robust results, achieving peak performance at ∼2.8k steps on NYT and ∼2.6k steps on Guardian. This configuration effectively balances exploration against stability. Visual analysis reveals that while larger groups ($G = 12$) facilitate rapid initial convergence, they suffer from training instability and performance fluctuations in later stages. Conversely, smaller groups ($G = 4$) restrict the exploration space, leading to suboptimal convergence and consistently lower accuracy.

4.3 Discussion

Efficiency Discussion. As shown in Tab. 5, with only 376M parameters, REFORM is significantly more compact than FKA-Owl (6.7B) and MMD-Agent (34B). Crucially, our dual-decoder design enables a Fast Mode (13.17 p/s) that bypasses the reasoning branch to achieve real-time screening speeds comparable to AMD. Since the reasoning and answer branches operate independently and in parallel, Fast Mode incurs zero accuracy loss compared to the Explainable Mode, because the answer decoder does not depend on the generated rationale at inference time. Thus, Fast Mode only removes the overhead of rationale generation, while leaving the final prediction unchanged.

Faithfulness of Teacher Rationales. To assess whether the distilled rationales are grounded in concrete multimodal evidence rather than merely restating manipulation labels, we conduct a blinded human audit on 350 ROM samples. As reported in Table 6a, the rationales recover 83.7% of the ground-truth visual evidence and 82.2% of the ground-truth textual evidence. These results indicate that the teacher rationales are generally evidence-grounded rather than template-like label paraphrases. Al-

Table 6: Analysis on rationale faithfulness (a) and teacher robustness (b).

Metric	Score (%)
GT visual evidence recall	83.7
GT textual evidence recall	82.2

(a) Blinded human audit on ROM.

Model	ACC	mAP	mIoU
REFORM (InternVL3.5-30B)	**81.52**	**67.75**	**81.46**
REFORM (Qwen2.5-VL-3B)	80.68 (-0.84)	66.29 (-1.46)	81.13 (-0.33)

(b) Robustness to teacher quality on Guardian setting of ROM.

though there is still room for improvement in rationale faithfulness, REFORM already yields substantial performance gains, supporting the effectiveness of our reasoning-driven learning paradigm.

Robustness to Teacher Quality. To examine whether REFORM's gains depend on a high-capacity annotator, we replace the original InternVL3.5-30B (Wang et al., 2025d) teacher with a much smaller Qwen2.5-VL-3B (Bai et al., 2025) model on the Guardian setting of ROM. As shown in Table 6b, performance drops remain small, with decreases of only 0.84 in ACC, 1.46 in mAP, and 0.33 in mIoU. These results suggest that the effectiveness of REFORM is not solely driven by a powerful teacher model, but is largely preserved even when the reasoning supervision is distilled from a smaller annotator.

5 Conclusion

In this paper, we address the generalization challenge by shifting from result-oriented supervision to explicit forensic reasoning. We propose REFORM, which integrates cognitive priming and a dual-decoder architecture, optimized via a GRPO-based curriculum to align judgment with logical evidence. Additionally, we contribute ROM, a large-scale benchmark expanding the scope to scene-level synthesis with reasoning annotations. Experiments demonstrate that REFORM significantly outperforms SOTA methods in cross-domain and zero-shot settings, establishing a robust paradigm for interpretable forensics.

6 Limitations

Despite the superior performance of REFORM, it still has several limitations that merit future study. **Residual Dependence on Distilled Rationales.** REFORM relies on distilled rationales during training. Although our analyses show that these rationales are generally evidence-grounded and that REFORM is reasonably robust to teacher quality, the final performance can still be affected by the faithfulness of the distilled supervision. Since we do not explicitly optimize rationale quality in this work, further improving it may lead to stronger forensic reasoning and better overall performance. **Inference Latency in Reasoning Mode.** While our Fast Mode is suitable for real-time screening, the Explainable Mode requires auto-regressive rationale generation, which increases computational cost (1.03 pairs/sec). Future work could explore more efficient rationale generation strategies.

7 Ethical Considerations

This work adheres to the ACL Code of Ethics. The ROM dataset and associated analyses were created solely to support research on detecting and grounding multimodal manipulations. We recognize that assembling realistic synthetic examples entails dual-use risks: the same materials and procedures could be misused to produce deceptive content. To minimize harm, we adopt a harm-minimizing, controlled-release approach: we will not publish the generation pipeline, detailed prompts, or prompt–response pairs to prevent their exploitation by adversaries for generating harmful content; public distribution is limited to vetted, research-only access under a signed Data Usage Agreement (DUA); distributed images will carry conspicuous visual watermarks and standardized metadata tags; high-fidelity originals and sensitive metadata will be withheld; images of minors and clearly sensitive contemporary conflict content have been excluded; and we reserve the right to revoke access on evidence of misuse. Full technical and procedural details of these safeguards are documented in the dataset README file.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62302140, and the National Key Research and Development Program of China under Grant No. 2023YFC3321600. The authors also gratefully acknowledge the support from the Guangdong Basic and Applied Basic Research Foundation (2025A1515012281), the Nanjing Municipal Science and Technology Bureau (202401035), and the University of Macau (MYRG-GRG2024-00077-FST-UMDF).

References

01.AI. 2024. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652v3*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xiao Bin, Wu Haiping, Xu Weijian, Dai Xiyang, Hu Houdong, Lu Yumao, Zeng Michael, Liu Ce, and Yuan Lu. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *ICML*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv Preprint arXiv:2507.06261*.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *ICLR*.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. 2022. Davit: Dual attention vision transformers. In *ECCV*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Jing Huang, Zhiya Tan, Shutao Gong, Fanwei Zeng, Joey Tianyi Zhou, Changtao Miao, Huazhe Tan, Weibin Yao, and Jianshu Li. 2025a. Lav-cot: Language-aware visual cot with multi-aspect reward optimization for real-world multilingual vqa. *arXiv preprint arXiv:2509.10026*.

Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, and Guangliang Cheng. 2025b. So-fake: Benchmarking and explaining social media image forgery detection. *arXiv preprint arXiv:2505.13379*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *EMNLP Findings*. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. 2025. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*.

Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. 2024. Generating attribution reports for manipulated facial images: A dataset and baseline. *arXiv preprint arXiv:2412.19685*.

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

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *ACM MM*.

Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *ICLR*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *ICLR*.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. 2025. S^2R: Teaching LLMs to self-verify

and self-correct via reinforcement learning. In *ACL*. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *EMNLP Findings*. Association for Computational Linguistics.

Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. In *EMNLP Findings*. Association for Computational Linguistics.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *CVPR*.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024a. Detecting and grounding multi-modal media manipulation and beyond. *TPAMI*, 46(8):5556–5574.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jinjie Shen, Yaxiong Wang, Lechao Cheng, Nan Pu, and Zhun Zhong. 2025. Beyond artificial misalignment: Detecting and grounding semantic-coordinated multimodal manipulations. In *ACM MM*.

StabilityAI. 2023. Introducing stable diffusion 3.5. Accessed: 2025-08-24.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Junxi Wang, Jize Liu, Na Zhang, and Yaxiong Wang. 2025a. Consistency-aware fake videos detection on short video platforms. In *International Conference on Intelligent Computing*, pages 200–210. Springer.

JunXi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. 2025b. FakeSV-VLM: Taming VLM for detecting fake short-video news via progressive mixture-of-experts adapter. In *EMNLP Findings*. Association for Computational Linguistics.

Qingyan Wang, Lianwei Wu, Yuanxia Zeng, Linyong Wang, Kang Wang, Yaxiong Wang, and Chao Gao. 2025c. Cross-modal consistency reasoning with large language models for short video-based fake news detection. In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, pages 37–45.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025d. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Zehong Yan, Peng Qi, Wynne Hsu, and Mong-Li Lee. 2025. TRUST-VL: An explainable news assistant for general multimodal misinformation detection. In *EMNLP*. Association for Computational Linguistics.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025a. Improve vision language model chain-of-thought reasoning. In *ACL*.

Yuchen Zhang, Yaxiong Wang, Yujiao Wu, Lianwei Wu, and Li Zhu. 2025b. The coherence trap: When mllm-crafted narratives exploit manipulated visual contexts. *arXiv preprint arXiv:2505.17476*.

Zhenxing Zhang, Yaxiong Wang, Lechao Cheng, Zhun Zhong, Dan Guo, and Meng Wang. 2025c. Asap: Advancing semantic alignment promotes multi-modal manipulation detecting and grounding. In *CVPR*.

Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, and Dong Yu. 2025. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*.

Cultivating Forensic Reasoning for Generalizable Multimodal Manipulation Detection

Supplementary Materials

A Experimental Setup

Implementation Details. All experiments are implemented in PyTorch and conducted on NVIDIA GeForce RTX 4090 GPUs using Distributed Data Parallel (DDP). The image encoder $\mathcal{E}_v$ adopts the DaViT architecture (Ding et al., 2022). Both the Multimodal Encoder $\mathcal{E}_m$ and the Cognitive Priming Encoder $\mathcal{E}_p$ are initialized with the pre-trained weights from Florence-2-B (Bin et al., 2023). Similarly, the Reason Decoder $\mathcal{D}_r$ and Answer Decoder $\mathcal{D}_a$ are initialized using the Florence-2-B decoder weights. The semantic alignment verifier V_ϕ is built upon the TinyBERT-4L-312D architecture (Jiao et al., 2020), comprising 4 transformer layers with a hidden dimension of 312. It employs a dual-head design—targeting 5-class image and 2-class text classification—where each head consists of a dense projection layer, GELU activation, and a dropout rate of 0.2. The input sequence length for V_ϕ is truncated to 320 tokens.

Parameter Settings. Input images are resized to 224×224 and augmented with random horizontal flipping. For the supervised training phases (Stage 1 and Stage 2), we set the per-GPU batch size to 5 and train the model for 4 and 12 epochs, respectively. Optimization is performed using AdamW (Loshchilov and Hutter, 2017) with an initial learning rate of 1×10^{-7} and a weight decay of 0.01. We employ a cosine learning rate scheduler with a linear warm-up: the learning rate increases to 1×10^{-6} over the first 1,000 steps and subsequently decays to 1×10^{-7}. In Stage 3, initialized with the best Stage 2 checkpoint, we set the generation group size to $G = 8$ to enhance exploration and optimization stability. The model is trained with a global batch size of 32 and a maximum sequence length of 1024. We utilize BF16 precision and Flash Attention 2 (Dao, 2024) for efficiency.

A.1 Baseline Settings

We adapt six state-of-the-art multi-modal methods to the ROM setting for comparison. These methods encompass two conventional multi-modal manipulation detection models, two Multimodal Large Language Model (MLLM) -based detection frameworks, one MLLM-based agentic manipulation detection framework, and one general multi-modal learning approach:

- **HAMMER** (Shao et al., 2023) is a pioneering model for multi-modal manipulation detection and grounding. It employs two unimodal encoders to extract visual and textual forgery features, which are then aligned through contrastive learning. Following this, a multi-branch transformer architecture with two specialized decoders is utilized for manipulation detection and grounding.

- **HAMMER++** (Shao et al., 2024a) is a more powerful model that builds upon HAMMER by integrating contrastive learning from both global and local perspectives to capture fine-grained inconsistencies.

- **FKA-Owl** (Liu et al., 2024) is a detection model designed on MLLM, demonstrating outstanding cross-domain performance. Since FKA-Owl does not natively support fine-grained classification tasks, we fine-tuned it using the same prompts as those used for REFORM to enable comparable evaluation.

- **AMD** (Zhang et al., 2025b) is a unified framework built upon MLLM designed for the MLLM-driven semantic-aligned DGM4 task. It introduces an Artifact Pre-perception Encoding module to capture manipulation traces into learnable artifact tokens and utilizes Manipulation-Oriented Reasoning to generate grounded detection results via a sequence-to-sequence format.

- **MMD-Agent** (Liu et al., 2025) is a training-free, MLLM-based agentic framework. It decomposes the detection task into sequential sub-tasks: text-based fact-checking (retrieving external knowledge), visual manipulation analysis, and cross-modal consistency verification. To adapt this pipeline for the grounding task in ROM, we modified its visual analysis prompt to explicitly request bounding box coordinates for manipulated regions.

- **ViLT** (Kim et al., 2021) serves as the general multi-modal learning baseline. It is a representative single-stream method where cross-modal interaction layers operate on the concatenation of image and text inputs. For adaptation to the forgery detection task, we add classification and detection heads to the corresponding outputs of the model.

A.2 Evaluation Metrics

To comprehensively evaluate our proposed ROM, we follow the rigorous evaluation protocols and metrics outlined in (Shao et al., 2023) for all manipulation detection and grounding tasks. The detailed evaluation setup is organized as follows:

- **Binary Classification. Accuracy (ACC)** is adopted as the evaluation metric to measure the correctness of real/fake news classification results.
- **Multi-Label Classification.** For multi-label classification tasks, we employ the **mean Average Precision (mAP)**, which measures the per-class average precision and then takes the arithmetic mean across all manipulation types. This macro-averaged mAP provides a comprehensive evaluation of the model's overall performance across different manipulation types.
- **Manipulated Image Bounding Box Grounding.** To evaluate the precision of predicted manipulated bounding boxes, we calculate the **mean Intersection over Union (mIoU)** between the ground-truth and predicted coordinates for all testing samples. This metric quantifies the spatial overlap between detected regions and actual manipulated areas, reflecting the localization accuracy of the model.
- **Manipulated Text Token Grounding.** In the DGM4 benchmark, an additional task of manipulated text token grounding is included. For this task, **Precision** (P_{tok}) is used as the evaluation metric to measure the accuracy of identifying manipulated text tokens within input sequences.

This standardized evaluation framework ensures a systematic and comparative assessment of ROM across diverse manipulation scenarios, aligning with both general detection tasks and benchmark-specific requirements.

A.3 Task-Specific Adaptation for DGM4

To adapt REFORM for fine-grained fake word detection on DGM4, we introduce a Token Precision Reward ($\mathcal{R}_{\text{tok}}$). Let $\hat{y}_{tok}$ denote the predicted manipulated words and y_{tok} be the ground truth token labels. The reward calculates the token-level consistency:

$$\mathcal{R}_{\text{tok}} = \text{ACC}(\text{Tokenize}(\hat{y}_{tok}), y_{tok}), \quad (13)$$

where, Tokenize$(\cdot)$ performs text normalization (lowercasing and punctuation removal) and aligns the word-level predictions with the caption's token grid to generate a binary prediction mask. ACC$(\cdot)$ computes the token-wise accuracy between this mask and the ground truth. Crucially, for pristine samples, this metric enforces strict hallucination suppression by assigning a reward of 1.0 only if the model predicts "none" or an empty set. Consequently, the final total Reward when training REFORM on DGM4 is formulated as:

$$\mathcal{R}_{\text{DGM4}} = \mathcal{R}_c + \mathcal{R}_a + \mathcal{R}_g + \mathcal{R}_f + \mathcal{R}_{\text{tok}}. \quad (14)$$

The training settings of other stages in REFORM are consistent with the main paper.

B Prompt Paradigm

In this section, we present the specific prompt templates designed for our proposed framework, the data distillation process, and the baseline comparisons. These templates ensure consistent task formulation across different experimental settings.

B.1 Prompt for REFORM

To enable fine-grained manipulation detection and localization, we design a structured prompt for the REFORM model. As shown in Fig. 5, the prompt concatenates system instructions, the news caption, and a specific set of ten options covering various manipulation types (e.g., face swap, text rewriting). Crucially, the task instruction explicitly requires the model to append the manipulated face's bounding box coordinates to the selected option if a face manipulation is detected.

B.2 Prompt for Reasoning Data Distillation

To equip our model with explicit reasoning capabilities, we distill knowledge from a powerful VLM, InternVL3.5-30B (Wang et al., 2025d). Fig. 6 illustrates the template used for reasoning generation. In this process, the ground-truth manipulation

Prompt Template for REFORM

[System Instruction]
<image> The following are multiple choice questions about fake news detection.

[Input Data]
The text caption of news is: *{News Caption}*

[Question & Options]
The image and text should not be manipulated. Question: Is there any manipulation in the image or text of this news?
A. No.
B. Image: Face swap; Text: No.
C. Image: Face attribute; Text: No.
D. Image: Whole generated; Text: No.
E. Image: Inpainted background; Text: No.
F. Image: Face swap; Text: Fully rewritten.
G. Image: Face attribute; Text: Fully rewritten.
H. Image: Whole generated; Text: Fully rewritten.
I. Image: Inpainted background; Text: Fully rewritten.
J. Image: No; Text: Fully rewritten.

[Task Instruction]
If the face is manipulated, locate the manipulated face in the image and append the results to your selected option.
The answer is: The answer is:

Figure 5: The prompt template for REFORM. The prompt strictly concatenates the system instruction, caption, options, and the localization instruction.

information is injected into the context, and the teacher model is instructed to generate a coherent, factual rationale explaining why the news falls into the target category, focusing on visual and textual evidence.

B.3 Prompt for General-purpose Model

For zero-shot comparisons with general-purpose MLLMs, we adapt the prompt to ensure parsing stability. As depicted in Fig. 7, while the input data and options remain consistent with the REFORM setting, we modify the task instruction to request normalized coordinates (relative positioning from 0 to 1). Furthermore, explicit constraints are added to suppress intermediate reasoning and force the model to output only the final option and coordinates.

B.4 Prompt for MMD-Agent

We also evaluate the MMD-Agent baseline. Since the original MMD-Agent pipeline does not natively support grounding for specific manipulation types, we modified the prompt in Stage 2 (Image Modality) to include localization instructions. Fig. 8 shows the complete workflow, where the agent sequentially performs text-based fact-checking, image manipulation detection (with our added bounding box requirement), and cross-modal consistency verification.

Prompt Template for Reasoning Generation

[System Instruction]
You are a fake news analysis expert. You will be provided with a news item in the form of an image-text pair.

[Ground Truth Context (Dynamic)]
It is known that *{Manipulation Description}* (e.g., "the image has been manipulated using face swap", "the image and text are original", etc.).

[Task Instruction]
Without assuming any prior knowledge about the authenticity or manipulation of this image-text news, analyze it carefully and reason only on the visual and textual evidence.

[Input Data]
The news text is: *{News Caption}*

[Reasoning Goal]
Finally, summarize your reasoning to justify why this news can be considered that *{Target Label Description}*.

[Constraints]
Keep your reasoning coherent, factual, and concise (*{Min Keywords}-{Max Keywords}* words). Write your response as plain continuous text, without any Markdown symbols, lists, or bullet points.

Figure 6: The prompt template used to generate reasoning chains. The *{Italicized Text}* represents dynamic content filled based on the ground-truth labels and input data.

Prompt Template for General-purpose Model

[System Instruction]
Same as REFORM
[Input Data]
Same as REFORM
[Question & Options]
Same as REFORM

[Task Instruction]
If face manipulation, use rectangular box coordinates in the format of [x1,y1,x2,y2], where the top-left vertex of the image is defined as (0,0) and the bottom-right vertex as (1,1) for relative positioning, and append the results to the option you have selected.
[Constraints]
Please think carefully before giving the finnal answer, but you don't have to output the reasoning process. ONLY output the final answer in format: [Option + Coordinates (if applicable).]

Figure 7: The prompt template used for General-purpose Model. Note the change in coordinate format to normalized [x1,y1,x2,y2].

Prompt Templates for MMD-Agent Workflow

[Stage 1: Fact-Checking (Text Modality)]
Given a news caption, news caption is: *{News Caption}*
Determine if there is credible objective evidence that SUPPORTS or REFUTES the news caption. Please follow the instructions below:
Thought 1: You need to find the key entity noun in the news caption. The key entity noun could be person or object or location or event, etc.
Action 1: Search [key entity noun].
Observation: *{External Knowledge / Wiki Results}*
Thought 2: According to Observation and other credible objective evidence, please analysis there is any objective fact that SUPPORTS or REFUTES the news caption, or if there is NOT ENOUGH INFORMATION. Analysis is: [Analysis].
Action 2: Draw the conclusion based on the analysis in the thought 2: if there is any credible objective evidence refuting the news caption, please answer in the form: 'Finish[TEXT REFUTES].'. If no, please answer in the form: 'Finish[TEXT SUPPORTS].'

[Stage 2: Manipulation Detection (Image Modality)]
According to the given news image, determine if the image is manipulated and identify the type of manipulation.
Classification categories: 'Real', 'Face Swap', 'Face Expression', 'Background Swap', 'Full AI Generated'.
Please follow the instructions below:
Thought 1: Analyze the image for visual inconsistencies such as unnatural lighting, blurred edges, distorted faces, or inconsistent background noise.
Observation: [Fact-conflicting Description]
Action 1: Draw the conclusion based on the observation.
- If Real: 'Finish[Real]'.
- If Face Swap/Expression: 'Finish[Face Swap, BBox: [x1, y1, x2, y2]]' (normalized 0.0-1.0).
- If Background/Full AI: 'Finish[Background Swap]' or 'Finish[Full AI Generated]'.

[Stage 3: Cross-Modal Consistency (Multimodal)]
Given a multimodal misinformation, it contains both news caption and news image. News caption is: *{News Caption}*
Determine if the news caption matches the content news image. You should answer in the following forms: 'Finish[MATCH].' or 'Finish[MISMATCH].'. Please follow the instructions below:
IMAGE DESCRIPTION: *{Image Description from Stage 2}*
Draw the conclusion: Based on the [IMAGE DESCRIPTION] of the news image, does the news caption match the content of news image? If yes, please answer in the form: 'Finish[MATCH].'. If no, please answer in the form: 'Finish[MISMATCH].'.

Figure 8: The MMD-Agent prompt workflow. The agent sequentially performs Fact-Checking (utilizing external knowledge), Image Manipulation Detection (providing bounding boxes for faces), and Consistency Verification, before aggregating the status to determine the final authenticity.