TextShield-R1: Reinforced Reasoning for Tampered Text Detection

Chenfan Qu[1, 3], **Yiwu Zhong**[2], **Jian Liu**[3 *], **Xuekang Zhu**[3], **Bohan Yu**[3], **Lianwen Jin**[1 *]

[1]South China University of Technology
[2] Peking University
[3] Ant Group
202221012612@mail.scut.edu.cn, rex.lj@antgroup.com, eelwjin@scut.edu.cn

Abstract

The growing prevalence of tampered images poses serious security threats, highlighting the urgent need for reliable detection methods. Multimodal large language models (MLLMs) demonstrate strong potential in analyzing tampered images and generating interpretations. However, they still struggle with identifying micro-level artifacts, exhibit low accuracy in localizing tampered text regions, and heavily rely on expensive annotations for forgery interpretation. To this end, we introduce TextShield-R1, the first reinforcement learning based MLLM solution for tampered text detection and reasoning. Specifically, our approach introduces Forensic Continual Pre-training, an easy-to-hard curriculum that well prepares the MLLM for tampered text detection by harnessing the large-scale cheap data from natural image forensic and OCR tasks. During fine-tuning, we perform Group Relative Policy Optimization with novel reward functions to reduce annotation dependency and improve reasoning capabilities. At inference time, we enhance localization accuracy via OCR Rectification, a method that leverages the MLLM's strong text recognition abilities to refine its predictions. Furthermore, to support rigorous evaluation, we introduce the Text Forensics Reasoning (TFR) benchmark, comprising over 45k real and tampered images across 16 languages, 10 tampering techniques, and diverse domains. Rich reasoning-style annotations are included, allowing for comprehensive assessment. Our TFR benchmark simultaneously addresses seven major limitations of existing benchmarks and enables robust evaluation under cross-style, cross-method, and cross-language conditions. Extensive experiments demonstrate that TextShield-R1 significantly advances the state of the art in interpretable tampered text detection.

Code — https://github.com/qcf-568/TextShield
Datasets — https://github.com/qcf-568/TextShield
Extended — https://github.com/qcf-568/TextShield

Introduction

The rapid advancement of image processing technologies has greatly lowered the barrier to creating tampered text images. Unfortunately, such forgeries are increasingly exploited for fraud, rumor dissemination, and other malicious purposes, posing significant security threats (Chen et al. 2024c). As a result, the reliable detection of tampered text has become a pressing research topic (Shao et al. 2024).

Recently, multimodal large language models (MLLMs) have demonstrated impressive human-like capabilities in perception and reasoning (Liu et al. 2025), making them highly promising for multiple tasks (Lan et al. 2024). In the context of tampered text detection, MLLMs can analyze visual and semantic artifacts while also generating textual justifications for their predictions, thereby enhancing the interpretability and trustworthiness of their decisions. However, despite their potential, MLLMs still face several critical limitations that hinder their effectiveness in this domain.

First, **inadequate task alignment**. Existing base MLLMs are typically pretrained on macro perception tasks centered around high-level semantics, such as image captioning and object recognition. In contrast, tampered text detection demands a micro-level perception to discern semantic-agnostic artifacts. This significant discrepancy makes the task overly challenging for MLLMs, often leading to confusion and overfitting during fine-tuning on tampered texts.

Second, **heavy annotation dependence**. Most current MLLMs depend heavily on costly forgery interpretation annotations, which are typically obtained through expensive closed-source models such as GPT-4o. However, due to privacy concerns, many credential images (e.g., ID cards, contracts) with sensitive information are prohibited from external exposure. Moreover, since forgery artifacts are often unobvious, automatic annotation is error-prone and demands extensive manual cleaning. These challenges hinder large-scale training on real-world data. Even if full annotations are produced through labor-intensive efforts, the "spoon-fed teaching" nature of supervised fine-tuning can compromise the MLLM's intrinsic reasoning and analytical capabilities.

Third, **poor localization accuracy**. MLLMs struggle to predict precise bounding boxes, especially for dense text. A naive solution involves integrating an extra traditional forgery localization model. However, this introduces additional latency. More importantly, such an approach can lead to misaligned predictions or excessive reliance on the localization model's biased and unrobust outputs, thereby undermining the MLLM's core aim of integration and universality.

To address these challenges and limitations, we propose **TextShield-R1** with innovations involving model pre-

* Corresponding Author

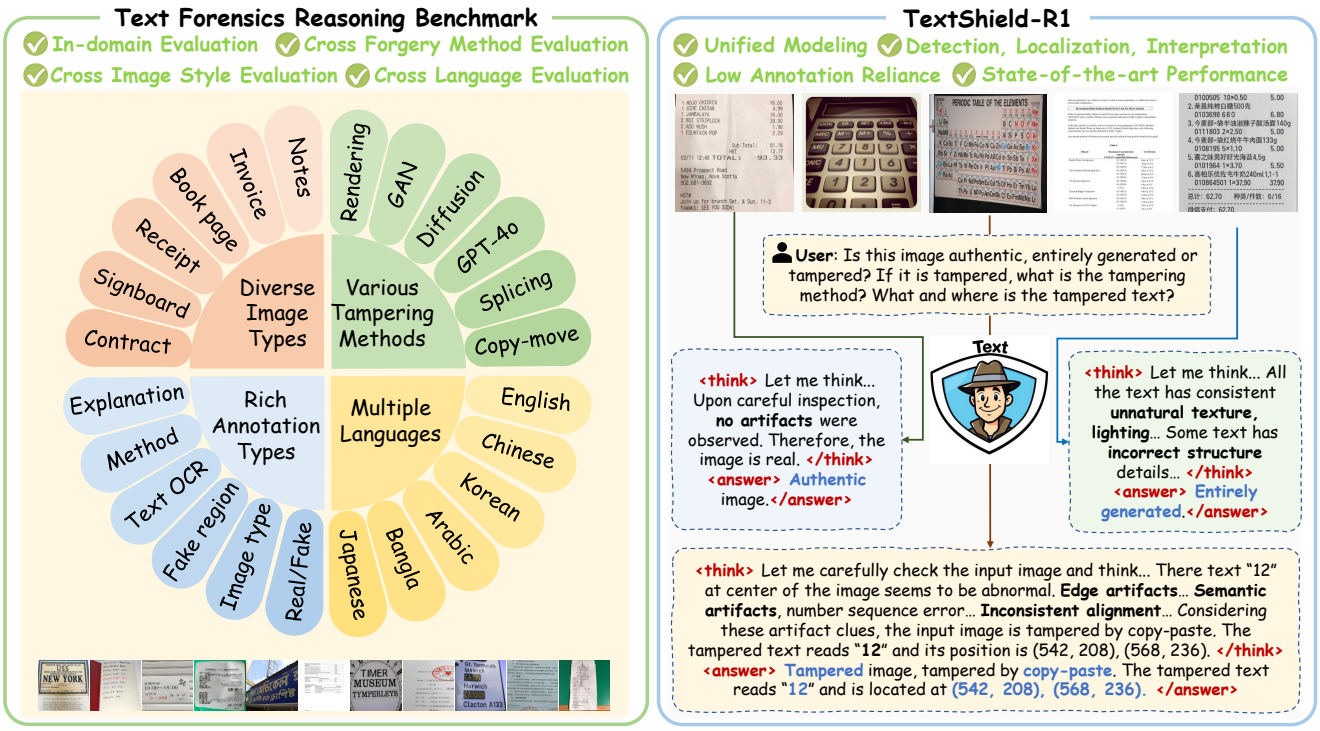

Figure 1: We introduce the Text Forensics Reasoning benchmark, which features a wide range of image domains, diverse and up-to-date tampering methods, rich annotations, various languages and comprehensive out-of-distribution evaluation settings. We also propose TextShield-R1, the first reinforcement learning based model for tampered text detection.

training, model fine-tuning, and model inference.

During pre-training, we propose Forensic Continual Pre-training to address task misalignment. It is an easy-to-hard curriculum that begins by training the MLLM to detect tampered natural objects, which typically exhibit prominent and detectable artifacts. By leveraging large-scale, high-quality natural image forgery datasets, this approach also enables the model to acquire robust and generalizable forensic features. We further introduce 3D Forensic Learning, which enhances supervision across three complementary dimensions. While pre-training on tampered natural objects equips the MLLM with essential forgery detection capabilities, it inevitably compromises their OCR ability. To mitigate this trade-off, we interleave the curriculum with an OCR reference grounding task, resulting in a model that is both forensically aware and OCR capable for downstream tampered text detection task.

During fine-tuning, to reduce reliance on costly artifact interpretation annotations, we introduce the first reinforcement learning approach tailored for tampered text detection. Specifically, the model is trained using Group Relative Policy Optimization (GRPO), guided by carefully crafted reward functions. This reinforcement learning approach not only mitigates the need for extensive annotations but also enhances the MLLM's reasoning capabilities.

During inference, to enhance the localization accuracy of tampered text, we introduce OCR Rectification. A task-specific OCR model first extracts both textual content and corresponding bounding box coordinates from the input image. Then, the MLLM predicts candidate tampered texts along with their associated bounding boxes. Each predicted box is refined by matching it to the OCR output based on content similarity and location proximity. If a suitable match is found, the OCR-derived box replaces the MLLM's original prediction. This approach effectively boosts localization performance by leveraging the MLLM's strong text recognition capabilities.

In addition to our proposed method, we introduce the Text Forensics Reasoning (TFR) benchmark, a comprehensive high-quality resource for tampered text detection. Our TFR addresses all the seven critical limitations of existing benchmarks: a **limited domain** that focuses solely on document or scene text; a **narrow scope** that excludes entirely generated text images; an **unbalanced ratio** lacking real samples for false-positive evaluation; **insufficient diversity** of tampering techniques; **outdated tampering methods**; **insufficient out-of-distribution** evaluation settings; and **incomplete annotations** missing textual artifact interpretations.

Our TFR benchmark addresses all these shortcomings, enabling more rigorous and realistic evaluation of tampered text detection methods across domains, languages, and tampering styles. Specifically, our benchmark comprises over 45k high-quality tampered text images and an equal number of real images with similar distributions. Notably, it is the first benchmark to comprehensively cover all three major text image types: documents, scene text, and ID-style

cards. It is also the first to include both locally tampered and fully generated text images, capturing a broader spectrum of real-world forgery scenarios. It is the first to support robust evaluation across image styles, forgery methods, and languages simultaneously. Furthermore, our TFR benchmark stands out for its state-of-the-art forgery quality. For instance, it is the first to include realistic fake text images generated using GPT-4o. To support deeper analysis, we also provide rich reasoning-style textual annotations, enabling more interpretable and LLM-compatible evaluations. We believe the TFR benchmark will serve as a valuable and foundational resource for advancing research in tampered text detection.

Extensive experiments conducted on the TFR benchmark and public benchmarks have validated our proposed method.

The main contributions of this paper are as follows:

- **TextShield-R1**, the first reinforcement learning method for unified tampered text detection, effectively optimized through our carefully designed reward functions.
- **Forensic Continual Pre-training**, an easy-to-hard curriculum that well prepares MLLM for tampered text detection by harnessing the cheap data from other tasks.
- **OCR Rectification**, an innovative method that advances MLLM's localization performance with OCR results.
- **Text Forensics Reasoning benchmark**, a comprehensive high-quality benchmark that addresses all the seven critical issues of previous benchmarks.

Related Works

Tampered Text Detection

Early works in tampered text detection relied on handcrafted features or rules, such as printer classification (Lampert, Mei, and Breuel 2006) or template matching (Ahmed and Shafait 2014). These methods are limited to specific layouts on scanned documents and do not work well on photographed text images (Wong et al. 2025). Some recent studies model tampered text detection as a semantic segmentation (Dong et al. 2024; Li et al. 2025c) or object detection (Wang et al. 2022a; Qu et al. 2025a) task. Despite progress, this prediction process remains a black box and produces unreliable results (Qu et al. 2024a). In addition, previous methods are tailored to a single image domain and cannot generalize across documents, scene texts, and ID cards. Some methods advanced the natural image forgery detection domain (Li et al. 2021, 2024b, 2025b; Qu et al. 2024b; Zhang et al. 2025; Zhu et al. 2025b; Su et al. 2025; Qu et al. 2025b; Ma et al. 2025; Yu et al. 2024) but do not work well on text images (Du et al. 2025; Li et al. 2025a).

MLLM-Driven Image Forensics

Recently, the rapid development of MLLM has accelerated progress in image forensics. FakeShield (Xu et al. 2024), ForgeryGPT (Li et al. 2024a) and SIDA (Huang et al. 2024) use MLLM to explain artifacts in natural images. M2F2-Det (Guo et al. 2025), TTFG (Sun et al. 2025) and AIGI-Holmes (Zhou et al. 2025) utilize MLLM to detect AI-generated face images. However, these methods depend heavily on forgery interpretation annotations. SoFake (Huang et al. 2025) leverages frozen SAM to detect AI-generated natural images. AvatarShield (Xu et al. 2025) harnesses temporal residual to detect AI-generated human videos. The above methods are designed for natural images. Due to the substantial differences between text forgeries and natural image forgeries (Luo et al. 2024), existing methods are not effective for tampered text detection.

Text Forensics Reasoning Benchmark

Motivation

As shown in Table 1, the construction of the Text Forensics Reasoning (TFR) is motivated by the seven crucial drawbacks of existing tampered text detection benchmarks:

• **Limited domain**: prior works are restricted to a single image domain (either document or scene text).

• **Narrow scope**: all previous benchmarks fail to include entirely generated text images, which represent an increasingly common real-world scenario.

• **Unbalanced ratio**: some benchmarks such as T-SROIE and DocTamper contain no real images, making it difficult to evaluate false positives effectively.

• **Insufficient diversity**: only a small number of tampering techniques are considered (only one or three techniques in most works), reducing the robustness of model evaluations.

• **Outdated quality**: Even the latest tampering method included in prior works is developed more than two years ago, lagging behind the rapid progress of forgery techniques.

• **Insufficient OOD evaluation**: existing benchmarks do not support thorough assessment of out-of-distribution (OOD) robustness, a crucial requirement for real-world deployment.

• **Incomplete annotations**: no public benchmark provides detailed textual reasoning annotations that analyze the visual and semantic artifacts of each tampered instance.

These crucial limitations significantly hinder the simulation of real-world scenarios and slow the development of practical text forensic methods.

Construction

We collected a diverse set of recent text images from document, scene text, and ID-style card domains. Image sources include the Internet and several public datasets. To create locally tampered text images, we applied copy-move, splicing, and rendering techniques using both manual efforts and meticulously designed automatic pipelines. Additionally, we included local text forgeries generated by a GAN model (SR-Net) and advanced diffusion models (DiffUTE (Chen et al. 2024a), TextDiffuser-2 (Chen et al. 2024b) and UDiffText (Zhao and Lian 2024)). Entirely generated images are produced by querying GPT-4o-image-1, TextDiffuser-2 (Chen et al. 2024b), AnyText-2 (Tuo, Geng, and Bo 2024), and Control-Net (Zhang, Rao, and Agrawala 2023) with varied prompts. Manual filtering is conducted to ensure quality.

In total, the TFR benchmark comprises 45,971 fake images paired with 45,514 corresponding real images. Sample images are displayed on the left side of Figure 2. To enable thorough OOD robustness evaluation, we defined three additional sub-sets beyond the common in-domain test set:

Dataset	Image Domain	Fake Region	Real Num.	Fake Num.	Method Num.	Lang Num.	Latest AIGC Method (#year)	OOD Evaluation			Forgery Explain.
								Style	Method	Lang.	
T-SROIE	Doc.	Local	0	920	1	1	SR-Net (2019)	×	×	×	×
DocTamper	Doc.	Local	0	170,000	3	2	-	✓	×	×	×
RTM	Doc.	Local	3000	6000	3	2	-	×	×	×	×
Tampered-IC13	S.T.	Local	233	229	1	1	SR-Net (2019)	×	×	×	×
OSTF	S.T.	Local	4412	1980	8	1	UDiffText (2023)	✓	✓	×	×
Ours	**Doc.+S.T. +Card**	**Local+ Global**	**45514**	45971	**10**	**16**	**GPT-4o (2025)**	✓	✓	✓	✓

Table 1: Comparison between public text forensic benchmarks, which include T-SROIE (Wang et al. 2022b), DocTamper (Qu et al. 2023), RTM (Luo et al. 2024), Tampered-IC13 (Wang et al. 2022a) and OSTF (Qu et al. 2025a). 'Doc.' denotes document, 'S.T.' denotes scene text, 'Num.' denotes number, 'Lang.' denotes language. 'OOD' includes out-of-distribution evaluation on unknown image styles, tampering methods and languages. 'Forgery Explain.' denotes textual forgery explanation annotation.

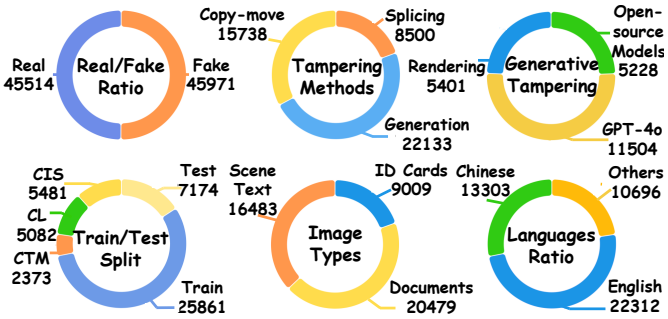

Figure 2: Representative samples (left) and data statistics (right) of the proposed Text Forensics Reasoning benchmark.

• Cross-Image-Style (CIS): Text images from sources not present in the training set.
• Cross-Tampering-Method (CTM): Text images tampered using three methods (TextDiffuser-2, SR-Net, Control-Net) excluded from the training set.
• Cross-Language (CL): Text images in 10 languages that differ from those in the training set.

Besides providing high-quality forgeries, we also constructed reasoning-style forgery interpretation annotations for all images. These textual annotations were obtained by querying GPT-4o to analyze both visual and semantic artifacts, followed by manual refinement to ensure quality. Basic statistics are provided on the right side of Figure 2

Highlights

As shown in Table 1, our TFR benchmark addresses all the seven major drawbacks of existing benchmarks and significantly outperforms them on multiple dimensions. For instance, our TFR features the most comprehensive coverage in terms of image domains (documents, scene text, and ID-style cards), forgery types (both local and global), tampering methods (10 techniques, including traditional methods and advanced AIGC methods), languages (16 types), out-of-distribution evaluation settings (3 distinct types), and annotation richness. This extensive scope establishes TFR as a valuable foundational resource for advancing text forensics.

TextShield-R1

In this section we present TextShield-R1, a novel method for tampered text detection and reasoning, distinguished by

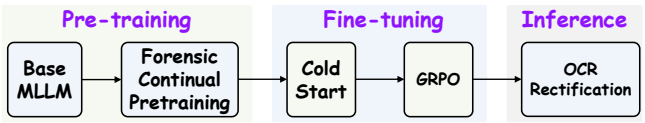

Figure 3: The overall pipeline of our TextShield-R1

advancements in both its training and inference pipelines. As illustrated in Figure 3, the training process begins with a **pre-training** stage involving Forensic Continual Pre-training on a base MLLM such as Qwen2.5-VL-7B. This pre-trained model then proceeds to a **fine-tuning** stage: initially with a small volume of fully annotated data to establish a cold start, followed by extensive fine-tuning with large-scale weakly annotated data using Group Relative Policy Optimization (GRPO). Here, "weakly annotated" implies that no artifact interpretation annotations are provided, requiring the model to reason about the artifact clues itself. In the **inference** stage, we utilize OCR Rectification to enhance localization accuracy. TextShield-R1's plug-and-play design requires no architectural modifications to the base MLLM, ensuring its broad applicability across diverse MLLMs.

Forensic Continual Pre-training

Existing base MLLMs are primarily pretrained to recognize semantics at a macro level. In contrast, tampering detection necessitates identifying semantic-agnostic artifacts at a micro-level. Besides the pre-training gap, the tampered text detection task is considerably challenging in nature, yet

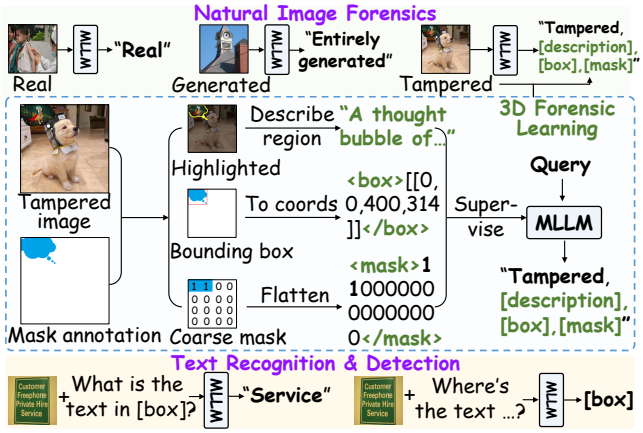

Figure 4: The Forensic Continual Pre-training pipeline. The MLLM is trained to distinguish between real, entirely generated, and locally tampered images. For locally tampered images, we introduce 3D Forensic Learning, which enhances supervision through three complementary dimensions. Additionally, we incorporate an OCR reference grounding task to prevent the forgetting of OCR-related knowledge.

high-quality training data remains costly and scarce. Consequently, directly fine-tuning an MLLM on this task inevitably leads to confusion and overfitting.

Inspired by the availability of large-scale, high-quality natural image forgeries, we propose to better prepare MLLMs for tampered text detection through continual pre-training on these datasets. Specifically, we pre-train the MLLM to classify whether an image is real, entirely generated, or locally tampered, as depicted in Figure 4. For locally tampered natural images, we introduce a 3D Forensic Learning approach to enhance supervision through task collaboration. In addition to tampering classification, we require the MLLM to output the description, bounding box coordinates and mask string of the tampered object. The tampered object description annotation is generated by inputting the tampered images and masks into the Describe Anything Model (Lian et al. 2025). The bounding box coordinates are obtained by calculating the minimum bounding boxes of the tampered regions. The mask string is generated by interpolating a mask annotation to 32x32 pixels and representing each mask as a 0/1 string, where '0' denotes a real region and '1' denotes a tampered region.

Through pre-training on natural image forgeries, the MLLM's artifacts perception and grounding capabilities can be notably advanced. However, training on these non-text images inevitably erodes MLLM's previously acquired OCR knowledge, which is indispensable for tampered text detection. To overcome the dilemma, we interleave an OCR reference-grounding task: given a real text image and a randomly chosen text instance, the model is either (a) provided with the bounding box and asked to output the text, or (b) provided with the text and asked to output the bounding box. This dual task preserves the model's OCR competence while reinforcing its forensic and grounding capabilities.

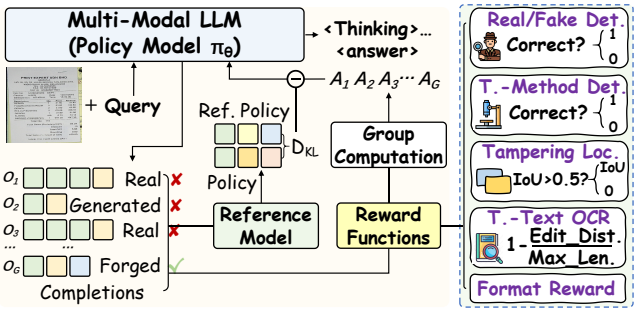

Figure 5: Under the GRPO framework, we optimize the model through five carefully designed reward functions.

Group Relative Policy Optimization

Existing MLLMs depend heavily on textual forgery interpretation annotations, the creation of which is both costly and potentially raises privacy concerns. Furthermore, the traditional "spoon-fed teaching" supervised fine-tuning can dampen an MLLM's native reasoning and analytical skills.

Recent advances in reinforcement learning have demonstrated immense potential for guiding and optimizing LLMs, particularly following the advent of Group Relative Policy Optimization (GRPO). Under the GRPO framework, we design a set of task-specific reward functions to better instruct the model. As depicted in Figure 5, our method incorporates five distinct rewards: Real/Fake classification reward, forgery method detection reward, tampering localization reward, tampered text OCR reward and format reward.

Real/Fake Classification Reward: we encourage accurate image-level three-way classification of real, entirely generated and locally tampered images. A reward of 1 is assigned for the correct classification, and 0 otherwise.

Forgery Method Detection Reward: For tampered image, we encourage the model to identify whether a fake region was created by copy-paste or generation. A reward of 1 is assigned for the correct identification, and 0 otherwise. Different forgery methods often leave distinct artifacts, and this reward helps the model achieve more in-depth analysis for improved interpretation and generalization.

Tampering Localization Reward: For tampered images, we encourage the model to accurately localize the tampered region. If the Intersection over Union (IoU) between the model's prediction and the ground truth label exceeds 0.5, the reward score is set to the IoU value; otherwise, it is 0.

Tampered Text OCR Reward: For tampered images, we encourage the model to accurately recognize the tampered text. We utilize the normed Levenshtein distance to quantify the similarity between the model's prediction and the ground truth string. The reward score is calculated as one minus the normed Levenshtein distance.

Format Reward: we encourage structured reasoning by rewarding outputs that embed reasoning within $\langle think \rangle ... \langle /think \rangle$ and answers within $\langle answer \rangle ... \langle /answer \rangle$ tags.

Through the proposed rewards, the model can effectively overcome its heavy reliance on textual annotations and fosters generalized analytical and reasoning skills.

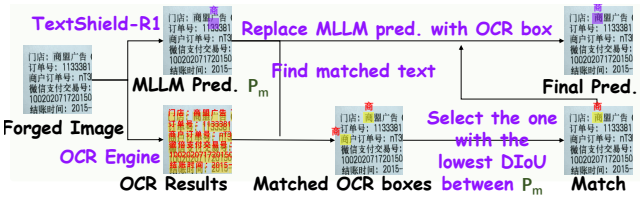

Figure 6: The proposed OCR Rectification pipeline, illustrating the case when multiple matched texts exist.

OCR Rectification

While MLLMs excel at text recognition, they generally struggle with predicting precise text box coordinates. Given that the text detection task is considerably simpler for task-specific OCR models, and highly accurate OCR engines are readily available, we propose to leverage the MLLMs' robust text recognition capabilities to refine their weak text localization predictions.

The proposed approach is illustrated in Figure 6. Given an image predicted to contain tampering, we obtain OCR results from an OCR engine. These OCR results include textual content and bounding box coordinates for each detected text instance. Subsequently, for each predicted tampered text, we search the OCR results for a matched text instance. The matched text is defined as the instance exhibiting the minimum Levenshtein distance with the model's prediction. If only one matched text is found, we directly replace the MLLM's localization prediction with the bounding box provided by the OCR engine for that match. If multiple matched texts exist, we select the instance that maximizes the Distance IoU with the MLLM's localization prediction. If no text instances satisfy the matching criterion (i.e., the normed Levenshtein distance between any OCR result and the predicted tampered text exceeds a fixed threshold of 0.2), we retain the MLLM's original localization prediction.

Through our OCR Rectification, the issue of inaccurate text localization in MLLMs is significantly mitigated.

Experiments

Implementation Details

We adopt the Qwen2.5-VL-7B as the base MLLM of our TextShield-R1. In the Forensic Continual Pre-training stage, 120k locally tampered natural images are collected from CASIAv1v2 (Dong, Wang, and Tan 2013), IMD20 (Novozamsky, Mahdian, and Saic 2020), NIST16 (Guan et al. 2019), MIML (Qu et al. 2024c) datasets; 120k entirely generated natural forgeries are collected from the Community Forensic (Park and Owens 2025) dataset; 60k images from the COCO (Lin et al. 2014) dataset and 60k images from the LAION (Schuhmann et al. 2021) dataset are collected as the authentic images. The authentic text images from the training set of TFR benchmark are used for the OCR reference grounding task. We pre-train our model on the collected data for one epoch. The model is LoRA (Hu et al. 2021) fine-tuned with LoRA rank 64 and is optimized by the AdamW optimizer with a learning rate decaying from 1e-4 to 0. In the fine-tuning stage, we train our model on the training set of the TFR benchmark. About 25% fully annotated data are used at first to establish a cold start, the rest of the images are fine-tuned under GRPO.

Comparison Study

We compare our method with both the official pre-trained versions and the TFR fine-tuned versions of MiniCPMV2.6 (Yao et al. 2024), Qwen2.5-VL-3B (Bai et al. 2025), Qwen2.5-VL-7B, InternVL3-2B (Zhu et al. 2025a) and InternVL3-8B. The results are shown in Table 2. Evidently, all the pre-trained models have low scores. This validates that the tampered text detection task is challenging and cannot be well solved by existing MLLMs. Additionally, our TextShield-R1 establishes a new state-of-the-art in the field of MLLM-based tampered text detection, which confirms the effectiveness of the proposed method in all the image-level classification, tampered text recognition, tampered text localization and artifacts reasoning tasks.

Ablation Study

Table 3 presents an ablation study of our proposed modules. Setting (1) is the Qwen2.5-VL-7B baseline, which includes none of our proposed modules, while setting (5) is our full TextShield-R1 model incorporating all of them. Removing Forensic Continual Pre-training (setting (2)) from the full model results in a significant performance degradation, with scores falling even below the baseline (1). This highlights the critical role of this pre-training stage, which establishes a foundational understanding of the complex text image forensics task. Without it, the model struggles to learn effectively and converge during the subsequent GRPO fine-tuning. The full model (5) performs comparably to setting (3) despite using textual forgery reasoning annotations for only a quarter of the training data. This result validates that our novel reward functions effectively enable the model to learn forgery reasoning from partially annotated datasets. Furthermore, by integrating our OCR Rectification method, the full model (5) outperforms setting (4) on the tampered text localization task across all four test sets. This confirms that OCR Rectification enhances localization performance by effectively leveraging the inherent OCR strengths of the MLLM.

Table 4 details the ablation study for our Forensic Continual Pre-training stage. Setting (1) is the baseline model without any continual pre-training. Setting (2), which involves pre-training solely on distinguishing between real, generated, and tampered images, improves image-level classification but leads to a significant drop in OCR and localization performance. This occurs because this narrow pre-training objective causes catastrophic forgetting of the model's inherent OCR and localization capabilities. Adding the 3D Forensic Learning task (setting (3)) notably improves localization performance, though the OCR score remains low due to the continued forgetting of text recognition knowledge. Our final pre-trained model (setting (5)) is achieved by adding the OCR Reference Grounding task, which makes the model both forensically aware and OCR-capable. Finally, the superior performance of setting (5) over setting (4) demonstrates that 3D Forensic Learning is essential, as it helps the model learn generalized features for forgery localization.

Method	Test set				CIS set				CTM set				CL set			
	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.
Official pre-trained base MLLMs without fine-tuning																
GPT4o	51.7	5.6	0.5	19.4	53.4	22.0	1.9	24.3	37.8	27.2	3.1	9.7	48.3	8.6	3.1	14.2
MiniCPM_V_2.6	30.4	1.6	0.0	3.2	31.4	4.8	0.0	3.2	26.2	7.0	0.0	2.5	30.9	0.4	0.0	1.9
InternVL3-2B	33.2	5.4	0.0	8.5	40.0	14.0	0.1	9.9	20.3	18.2	0.4	4.0	34.5	6.7	0.0	6.9
InternVL3-8B	40.4	9.3	0.2	17.9	47.0	20.9	0.8	20.3	25.2	31.5	1.7	8.8	45.1	10.3	0.5	18.1
Qwen2.5-VL-3B	46.2	1.8	0.1	9.5	48.4	5.9	0.3	15.3	42.2	7.7	0.2	5.6	47.1	1.5	0.2	7.9
Qwen2.5-VL-7B	42.6	6.4	0.1	9.5	49.9	19.4	0.4	17.6	34.0	21.0	0.6	4.5	50.1	11.1	0.2	10.4
MLLMs fine-tuned with full training set images																
MiniCPM_V_2.6	76.2	17.6	11.2	41.1	71.7	24.5	22.8	32.3	64.8	20.7	24.9	27.3	81.5	33.6	20.5	40.3
InternVL3-2B	75.4	18.1	10.3	40.6	68.5	23.1	21.4	32.0	62.5	18.7	26.2	25.0	80.2	31.7	21.0	39.5
InternVL3-8B	78.6	21.9	15.4	41.7	70.8	27.6	25.2	33.8	67.7	23.6	31.4	32.3	84.3	38.0	24.8	42.0
Qwen2.5-VL-3B	77.5	18.6	11.6	42.9	72.3	25.0	20.9	33.6	63.0	18.8	25.6	24.6	80.9	32.4	21.4	39.7
Qwen2.5-VL-7B	79.1	24.3	18.2	42.9	71.1	30.7	26.5	35.7	73.6	26.3	34.2	36.2	85.1	38.2	25.5	43.1
FakeShield	70.5	9.2	5.4	35.6	62.0	14.8	10.6	29.2	57.5	15.1	17.3	21.4	70.3	23.9	11.3	34.8
FakeShield*	79.1	24.3	7.6	42.8	71.1	30.5	15.0	35.6	73.6	26.3	21.8	36.2	85.1	38.1	15.6	42.9
SIDA	71.2	9.2	5.6	35.7	62.2	14.9	10.8	29.4	57.5	15.1	17.3	21.5	70.4	23.8	11.5	25.0
SIDA*	79.2	24.3	7.7	42.9	71.4	30.9	15.1	35.7	73.6	26.3	21.8	36.3	85.2	38.2	15.8	43.0
Ours	**88.1**	**47.6**	**57.8**	**58.8**	**72.9**	**62.1**	**61.0**	**56.5**	**88.8**	**45.6**	**68.3**	**51.2**	**85.5**	**39.0**	**40.6**	**46.2**

Table 2: Comparison experiments. 'Cls' denotes the real/generated/tampered task with the accuracy metric. 'OCR' denotes the tampered text recognition task with the OCR accuracy metric. 'Loc.' denotes the tampered text localization task with the IoU metric. 'Res.' denotes the forgery reasoning task, using the average score of cosine similarity, Rouge-L and BLEU as metric. 'FakeShield*', 'SIDA*' denote FakeShield (Xu et al. 2024) and SIDA (Huang et al. 2024) with the Qwen2.5-VL-7B as MLLM.

Num	Ablation	Test set				CIS set				CTM set				CL set			
		Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.
(1)	Baseline	79.1	24.3	18.2	42.9	71.1	30.7	26.5	35.7	73.6	26.3	34.2	36.2	85.1	38.2	25.5	43.1
(2)	w.o. FCP	75.8	21.9	12.7	39.0	68.4	25.0	20.9	30.6	66.3	23.7	25.5	30.1	83.9	38.5	26.0	41.8
(3)	w.o. GRPO	87.6	46.8	57.7	58.6	72.3	61.7	60.8	56.2	88.1	45.3	68.2	50.9	85.4	38.5	40.2	46.1
(4)	w.o. OCR Rect.	**88.1**	**47.6**	42.7	**58.8**	**72.9**	**62.1**	56.6	**56.5**	**88.8**	**45.6**	57.9	**51.2**	**85.5**	**39.0**	32.3	**46.2**
(5)	TextShield-R1	**88.1**	**47.6**	**57.8**	**58.8**	**72.9**	**62.1**	**61.0**	**56.5**	**88.8**	**45.6**	**68.3**	**51.2**	**85.5**	**39.0**	**40.6**	**46.2**

Table 3: Ablation study on the proposed modules. 'w.o.' denotes 'without'. 'FCP' denotes the Forensic Continual Pre-training approach. 'OCR Rect.' denotes the proposed OCR Rectification.

Num.	Ablations			Test set				CIS set				CTM set				CL set			
	Nat.	3D-FL	OCR	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.	Cls.	OCR	Loc.	Res.
(1)	×	×	×	75.8	21.9	12.7	39.0	68.4	25.0	20.9	30.6	66.3	23.7	25.5	30.1	83.9	38.5	26.0	41.8
(2)	✓	×	×	80.9	12.7	9.8	34.0	65.1	14.3	12.4	27.5	57.6	16.9	15.1	26.2	80.1	18.6	17.4	36.5
(3)	✓	✓	×	82.3	11.5	13.9	39.2	67.7	12.0	15.8	29.8	63.6	14.7	19.1	28.4	81.5	9.9	22.9	37.2
(4)	✓	×	✓	83.2	40.9	48.6	52.7	70.4	56.8	52.0	51.4	78.6	41.1	56.1	48.3	82.5	37.8	34.2	43.0
(5)	✓	✓	✓	**88.1**	**47.6**	**57.8**	**58.8**	**72.9**	**62.1**	**61.0**	**56.5**	**88.8**	**45.6**	**68.3**	**51.2**	**85.5**	**39.0**	**40.6**	**46.2**

Table 4: Ablation study on the proposed Forensic Continual Pre-training method. 'Nat.' denotes pre-training the model to identify whether a natural image is real/generated/tampered. '3D-FL' denotes the proposed 3D Forensic Learning approach. 'OCR' denotes including the OCR reference grounding task.

Conclusion

In this work, we introduced TextShield-R1, a novel framework that systematically addresses key challenges in MLLM-based tampered text detection. Our Forensic Continual Pre-training bridges the gap between general-purpose pre-training and fine-grained forensic analysis using an easy-to-hard curriculum. To reduce reliance on expensive annotations and foster deeper analytical skills, we pioneered a reinforcement learning approach, which guides the model with novel reward functions. Furthermore, our OCR Rectification method elegantly resolves poor localization accuracy by leveraging the MLLM's own powerful text recognition capabilities to refine its predictions. We also constructed the Text Forensics Reasoning (TFR) benchmark. This comprehensive resource remedies seven major deficiencies of prior datasets by incorporating diverse image domains, modern forgery techniques, and robust cross-domain, cross-method, and cross-language test settings. Extensive experiments validate that TextShield-R1 significantly advances the state of the art in detection accuracy, generalization, and interpretability. By tackling critical gaps in both methodology and evaluation, our work provides a robust foundation for future research into developing more reliable and trustworthy forensic AI systems.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (Grant No.:62476093) and Ant Group Research Intern Program.

References

Ahmed, A. G. H.; and Shafait, F. 2014. Forgery detection based on intrinsic document contents. In *2014 11th IAPR International Workshop on Document Analysis Systems*, 252–256. IEEE.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, H.; Xu, Z.; Gu, Z.; Li, Y.; Meng, C.; Zhu, H.; Wang, W.; et al. 2024a. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2024b. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, 386–402. Springer.

Chen, Z.; Chen, S.; Yao, T.; Sun, K.; Ding, S.; Lin, X.; Cao, L.; and Ji, R. 2024c. Enhancing Tampered Text Detection Through Frequency Feature Fusion and Decomposition. In *European Conference on Computer Vision*, 200–217. Springer.

Dong, J.; Wang, W.; and Tan, T. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 422–426.

Dong, R. L., Li; Ma, B.; Zhang, W.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; and Cheng, X. 2024. Robust Text Image Tampering Localization via Forgery Traces Enhancement and Multiscale Attention. *IEEE Transactions on Consumer Electronics*.

Du, B.; Zhu, X.; Ma, X.; Qu, C.; Feng, K.; Yang, Z.; Pun, C.-M.; Liu, J.; and Zhou, J. 2025. ForensicHub: A Unified Benchmark & Codebase for All-Domain Fake Image Detection and Localization. *arXiv preprint arXiv:2505.11003*.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 63–72. IEEE.

Guo, X.; Song, X.; Zhang, Y.; Liu, X.; and Liu, X. 2025. Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 105–116.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Z.; Hu, J.; Li, X.; He, Y.; Zhao, X.; Peng, B.; Wu, B.; Huang, X.; and Cheng, G. 2024. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. arXiv:2412.04292.

Huang, Z.; Li, T.; Li, X.; Wen, H.; He, Y.; Zhang, J.; Fei, H.; Yang, X.; Huang, X.; Peng, B.; et al. 2025. So-Fake: Benchmarking and Explaining Social Media Image Forgery Detection. *arXiv preprint arXiv:2505.18660*.

Lampert, C. H.; Mei, L.; and Breuel, T. M. 2006. Printing technique classification for document counterfeit detection. In *2006 International Conference on Computational Intelligence and Security*, volume 1, 639–644. IEEE.

Lan, M.; Chen, C.; Zhou, Y.; Xu, J.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2024. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*.

Li, J.; Zhang, F.; Zhu, J.; Sun, E.; Zhang, Q.; and Zha, Z.-J. 2024a. ForgeryGPT: Multimodal Large Language Model For Explainable Image Forgery Detection and Localization. arXiv:2410.10238.

Li, S.; Guo, Y.; Chen, S.; Li, B.; Lin, K.; Chen, C.; Li, H.; Yao, T.; and Ding, S. 2025a. DITL2: Dual-Stage Invariance Transfer Learning for Generalizable Document Image Tampering Localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 82–91. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024b. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Li, S.; Xing, Z.; Wang, H.; Hao, P.; Li, X.; Liu, Z.; and Zhu, L. 2025b. Toward Medical Deepfake Detection: A Comprehensive Dataset and Novel Method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 626–637. Springer.

Li, S.; Xu, S.; Ma, W.; and Zong, Q. 2021. Image manipulation localization using attentional cross-domain CNN features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 5614–5628.

Li, W.; Li, B.; Zheng, K.; Li, S.; and Li, H. 2025c. Document image forgery detection and localization in desensitization scenarios. *Signal Processing*, 110123.

Lian, L.; Ding, Y.; Ge, Y.; Liu, S.; Mao, H.; Li, B.; Pavone, M.; Liu, M.-Y.; Darrell, T.; Yala, A.; et al. 2025. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Luo, D.; Liu, Y.; Yang, R.; Liu, X.; Zeng, J.; Zhou, Y.; and Bai, X. 2024. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 110828.

Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.-M.; Lv, J.; et al. 2025. Imdl-benco: A

comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37: 134591–134613.

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*.

Park, J.; and Owens, A. 2025. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8245–8257.

Qu, C.; Liu, C.; Liu, Y.; Chen, X.; Peng, D.; Guo, F.; and Jin, L. 2023. Towards robust tampered text detection in document image: new dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5937–5946.

Qu, C.; Liu, J.; Chen, H.; Yu, B.; Liu, J.; Wang, W.; and Jin, L. 2024a. TextSleuth: Towards Explainable Tampered Text Detection. *arXiv preprint arXiv:2412.14816*.

Qu, C.; Zhong, Y.; Guo, F.; and Jin, L. 2024b. Omni-IML: Towards Unified Image Manipulation Localization. *arXiv preprint arXiv:2411.14823*.

Qu, C.; Zhong, Y.; Guo, F.; and Jin, L. 2025a. Revisiting tampered scene text detection in the era of generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 694–702.

Qu, C.; Zhong, Y.; Li, B.; and Jin, L. 2025b. Webly-Supervised Image Manipulation Localization via Category-Aware Auto-Annotation. *arXiv preprint arXiv:2508.20987*.

Qu, C.; Zhong, Y.; Liu, C.; Xu, G.; Peng, D.; Guo, F.; and Jin, L. 2024c. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shao, H.; Qian, Z.; Huang, K.; Wang, W.; Huang, X.; and Wang, Q. 2024. Delving into adversarial robustness on document tampering localization. In *European Conference on Computer Vision*, 290–306. Springer.

Su, L.; Ma, X.; Zhu, X.; Niu, C.; Lei, Z.; and Zhou, J.-Z. 2025. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7024–7032.

Sun, K.; Chen, S.; Yao, T.; Zhou, Z.; Ji, J.; Sun, X.; Lin, C.-W.; and Ji, R. 2025. Towards general visual-linguistic face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19576–19586.

Tuo, Y.; Geng, Y.; and Bo, L. 2024. AnyText2: Visual Text Generation and Editing With Customizable Attributes. arXiv:2411.15245.

Wang, Y.; Xie, H.; Xing, M.; Wang, J.; Zhu, S.; and Zhang, Y. 2022a. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, 215–232. Springer.

Wang, Y.; Zhang, B.; Xie, H.; and Zhang, Y. 2022b. Tampered text detection via rgb and frequency relationship modeling. *Chinese Journal of Network and Information Security*, 8(3): 29–40.

Wong, K.; Zhou, J.; Wu, H.; Si, Y.-W.; and Zhou, J. 2025. ADCD-Net: Robust Document Image Forgery Localization via Adaptive DCT Feature and Hierarchical Content Disentanglement. *arXiv preprint arXiv:2507.16397*.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2024. FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models. arXiv:2410.02761.

Xu, Z.; Zhang, X.; Zhou, X.; and Zhang, J. 2025. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yu, J.; Lu, D.; Shi, X.; Qu, C.; and Guo, F. 2024. Unified Face Attack Detection with Micro Disturbance and a Two-Stage Training Strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 960–969.

Zhang, L.; Li, S.; Ma, W.; and Zha, H. 2025. TrueMoE: Dual-Routing Mixture of Discriminative Experts for Synthetic Image Detection. *arXiv preprint arXiv:2509.15741*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhao, Y.; and Lian, Z. 2024. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European conference on computer vision*, 217–233. Springer.

Zhou, Z.; Luo, Y.; Wu, Y.; Sun, K.; Ji, J.; Yan, K.; Ding, S.; Sun, X.; Wu, Y.; and Ji, R. 2025. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models. *arXiv preprint arXiv:2507.02664*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; and Zhou, J.-Z. 2025b. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11022–11030.