TRACETOK: FINE-GRAINED TEXT TAMPERING LOCALIZATION VIA NATIVE MASK TOKEN GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text images are pervasive in financial, legal, and administrative records, where text tampering can undermine the credibility of the underlying information. Existing forensic multimodal large language models (MLLMs) can localize suspicious regions and provide natural-language analysis. However, they localize tampering through textual bounding boxes or tokens coupled to segmentation decoders, which are coarse for irregular shapes and fragmented regions not aligned with object-level semantic boundaries. Moreover, MLLM visual encoders are less sensitive to subtle manipulation traces than dedicated forensic experts. To address these limitations, we propose TraceTok, a forensic-enhanced MLLM that formulates fine-grained tampered text localization as native mask-token generation and jointly produces textual forensic responses and pixel-level tampering masks. The Tamper-Aware Adaptive-Length Mask Reconstruction module (TAMR) conditions mask reconstruction on image-side forensic features and uses Token Length Predictor (TLP)-guided retained-length selection, enabling compact token prefixes to preserve small and irregular tampered text structures. Evidence-Preserving Adaptive Grid Fusion (EAGF) aligns multistage expert features with the current visual-token grid before fusing and injecting them into the visual branch, thereby strengthening low-level forensic perception across varied input resolutions. On TFR-Test, CIS, and CTM, TraceTok attains 62.2 average Mask IoU, exceeding our reevaluated TextShield-R1 baseline by 12.7 points; on RTM it matches the strongest task-specific localizers, and its reconstructor reaches comparable or better F1 with fewer retained tokens than previous methods.

1 INTRODUCTION

Text manipulation is increasingly accessible (Chen et al., 2023; Tuo et al., 2024), yet text images remain central to financial, legal, and administrative workflows. Reliable tampered-text detection, pixel-level localization, and evidence-based explanation are therefore needed (Chen et al., 2024b).

Existing Tampered Text Localization (TTL) methods model RGB appearance, frequency artifacts, compression traces, and spatial text layout to localize edits (Wang et al., 2022; Qu et al., 2023; Luo et al., 2025; Chen et al., 2024b; Yi et al., 2026). Recent multimodal large language model (MLLM) methods connect visual evidence with natural-language analysis (Xu et al., 2025b; Huang et al., 2025a; Qu et al., 2024; Xu et al., 2025a; Qu et al., 2026). These methods use different localization schemes to specify suspected regions. LISA-style methods emit a special token, such as <SEG>, whose hidden representation an external decoder maps to a mask (Xu et al., 2025b; Huang et al., 2025a; Lai et al., 2024; Kirillov et al., 2023; Xia et al., 2024; Rasheed et al., 2024; Ren et al., 2024; Zhang et al., 2024b). TextSleuth and TextShield-R1 specify suspected regions with a textual bounding box (Qu et al., 2024; 2026). TVSIP uses an MLLM to identify suspicious text and predict bounding boxes, refines them with OCR, and fuses them with an expert model's mask to guide a subsequent MLLM interpreter (Xu et al., 2025a). These localization schemes lack a common output representation that the MLLM can generate alongside its forensic response and supervise directly with a target tampering mask. Figure 1 summarizes this distinction between localization schemes.

Despite this progress, current MLLM-based TTL methods face two limitations. (1) Indirect Spatial Grounding. In LISA-style methods, the localization token is supervised only indirectly, through the mask loss that the external decoder passes back, rather than by a mask-level target on the token itself.

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

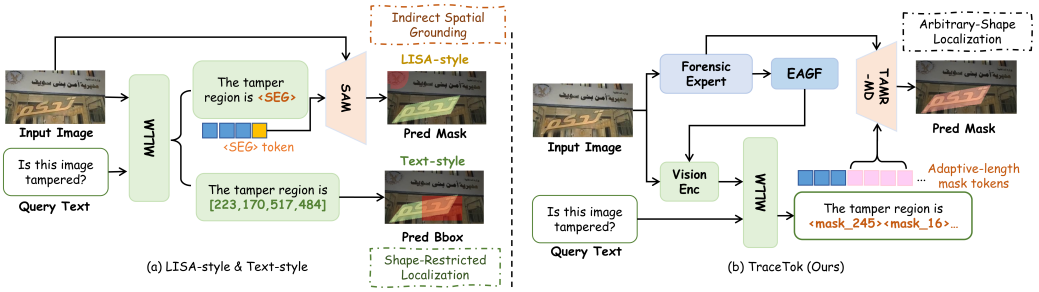


Figure 1: Overview of MLLM localization schemes and TraceTok. (a) LISA-style methods emit a segmentation token decoded by an external module, and Text-style methods return textual bounding boxes. (b) TraceTok generates adaptive-length mask tokens and reconstructs a pixel-level mask with TAMR, while EAGF injects aligned forensic evidence.

A tampered region is also not necessarily a semantically well-defined object, which makes precise localization difficult for the external decoder. Text-style methods specify suspicious regions through a textual bounding box rather than a mask, and such an output cannot faithfully represent tampered regions with arbitrary shapes. (2) Limited Sensitivity to Forensic Traces. LISA-style and Text-style methods primarily rely on semantic visual features from the MLLM's visual encoder. These features support high-level interpretation but are less sensitive to subtle manipulation traces captured by forensic expert models (Guillaro et al., 2023; Su et al., 2025). LoRA-based adaptation (Hu et al., 2022) alone may not provide the task-specific low-level evidence needed to improve this sensitivity, leaving the MLLM less responsive to forensic cues relevant to tampered-text localization.

We propose TraceTok, a forensic-enhanced generative multimodal large language model (MLLM) that generates textual forensic responses and performs pixel-level tampered-text localization within a single autoregressive process. To address the first limitation, we formulate localization as autoregressive mask-token generation, allowing the MLLM to produce a spatial mask representation alongside its textual response. For tampered-text regions that are often small, fragmented, and irregular along text boundaries, we further develop TAMR, an adaptive-length mask reconstruction module. TAMR reconstructs tampering masks from retained mask-token sequences conditioned on image-side forensic evidence, and uses TLP-guided retained-length selection to adapt the token budget to mask complexity, preserving key local structures under a limited budget. To address the second limitation, we augment the MLLM visual branch with complementary forensic expert features to provide low-level evidence beyond semantic visual representations. Because varying input resolutions change the spatial sizes of expert features, we introduce EAGF, an evidence-preserving adaptive grid fusion module that adaptively aligns multistage forensic expert features with the current MLLM visual-token grid, fuses their complementary evidence, and injects it into the visual branch through gated residual connections. As illustrated in Fig. 1, TraceTok addresses the limitations of Indirect Spatial Grounding and Limited Sensitivity to Forensic Traces using native mask-token generation and forensic feature fusion. It achieves an average Mask IoU of 62.2 on the three evaluated TFR subsets, improving over our reevaluated TextShield-R1 baseline by 12.7 points on average.

In summary, our main contributions are as follows:

- We propose TraceTok, a forensic-enhanced MLLM that unifies textual forensic responses and pixel-level tampered-text localization through autoregressive mask-token generation.
- We propose TAMR, an adaptive-length mask reconstruction module that improves the spatial precision of MLLM localization. It adapts the retained length of mask-token sequences to mask complexity, preserving character-level details and small foreground regions.
- We propose EAGF, an evidence-preserving adaptive grid fusion module that addresses the lack of low-level forensic perception in MLLMs. It injects complementary forensic expert features into the visual branch and aligns them with the input-dependent visual-token grid.
- Experiments on TFR and RTM show that TraceTok improves MLLM-based localization and mask reconstruction with fewer retained tokens, while remaining competitive with task-specific localizers.

2 Related Work

Tampered Text Localization. Early TTL methods use RGB appearance and frequency-domain cues to localize scene-text edits (Wu et al., 2019; Wang et al., 2022). OSTF broadened evaluation with diverse text-editing methods and open-set scene-text tampering (Qu et al., 2025). For document images, DocTamper introduced a large-scale benchmark and DTD, which uses compression artifacts for robust localization (Qu et al., 2023). FFDN fuses RGB and frequency information through visual and wavelet-like frequency enhancement (Chen et al., 2024b). RTM provides a manually tampered document dataset and a tampered-authentic contrastive learning strategy for realistic text manipulation (Luo et al., 2025). Other studies address spatial-noise cross-attention, adversarial robustness, forgery-trace enhancement and multiscale attention, cross-domain generalization, structured document priors, and frequency information aggregated around text regions, as well as DCT-based document-image cues (Song et al., 2025; Shao et al., 2024; Dong et al., 2024; Li et al., 2025a; Zheng et al., 2026; Li et al., 2025b; Yi et al., 2026; Wong et al., 2025; Li et al., 2026). These methods provide task-specific localization, but their outputs offer limited support for human inspection.

MLLMs for Image Forensics. MLLM-based image forensics links visual evidence with natural-language explanations. FakeShield and SIDA jointly address forgery detection, localization, and explanation, while ForgeryGPT uses expert models for interpretable forgery analysis (Xu et al., 2025b; Huang et al., 2025a; Zhang et al., 2024a; Sun et al., 2025; Huang et al., 2025b; Nguyen et al., 2024; Zhang et al., 2026). In TTL, TextSleuth provides an explainable benchmark and region-focused analysis. TVSIP pairs a low-level expert locator with an MLLM-based interpreter, and TextShield-R1 uses forensic continual pre-training, reinforcement learning, and OCR rectification (Qu et al., 2024; Xu et al., 2025a; Qu et al., 2026). These methods localize with textual bounding boxes or LISA-style special tokens linked to external segmentation decoders (Lai et al., 2024; Kirillov et al., 2023). AIGI-Holmes also uses forensic expert features in an MLLM-based visual pipeline (Zhou et al., 2025). HiMTok and ALToLLM encode masks as discrete token sequences an MLLM can generate autoregressively; ALToLLM further adjusts the sequence length to mask complexity (Wang et al., 2025b;a; Lan et al., 2025; Yu et al., 2024). ALToLLM learns this behavior through random-prefix reconstruction followed by TLP updating. TraceTok builds on these directions with tamper-aware adaptive-length mask reconstruction and visual-token-grid-aware evidence fusion.

3 Method

3.1 TraceTok Framework

TraceTok formulates TTL as an instruction-guided adaptive-length mask-token generation task. Given an image and a task instruction, the MLLM generates a textual forensic response and, when localization is required, an adaptive-length mask-token sequence (Sec. S4). TAMR decodes this sequence into a tampering mask with its mask de-tokenizer (TAMR-MD). Figure 2 summarizes the four-stage training pipeline of TraceTok (Sec. S1).

1) TAMR follows the mask-token reconstruction paradigm of ALToLLM: it learns to encode a tampering mask into a codebook sequence of mask tokens and decode the sequence back into a reconstructed mask. At this stage, a frozen forensic expert encoder supplies features that are fused with the mask-token representation before TAMR-MD, providing image-side forensic evidence for reconstruction. Randomly sampling the retained mask-token length during reconstruction provides the training signal for the Token Length Predictor (TLP) to initially learn the correspondence between mask complexity and retained length. 2) We use the TLP initialized in Stage 1 to predict a retained length for each mask and define a bounded sampling window around this prediction. This TLP-guided refinement makes the retained length follow mask complexity while concentrating useful information in earlier tokens; the expert features continue to condition TAMR-MD. 3) The MLLM is pre-trained with image-category detection and bidirectional text–mask grounding, establishing forensic-aware visual perception and text–mask correspondence. 4) We use the trained TAMR and forensic expert encoder during MLLM fine-tuning to enable joint textual response generation and adaptive-length mask-token localization. A ground-truth mask encoded by TAMR-MT is inserted into the target language sequence as mask-token supervision. The frozen features F_3 and F_4 from the forensic expert encoder condition TAMR-MD during mask reconstruction and enter the visual

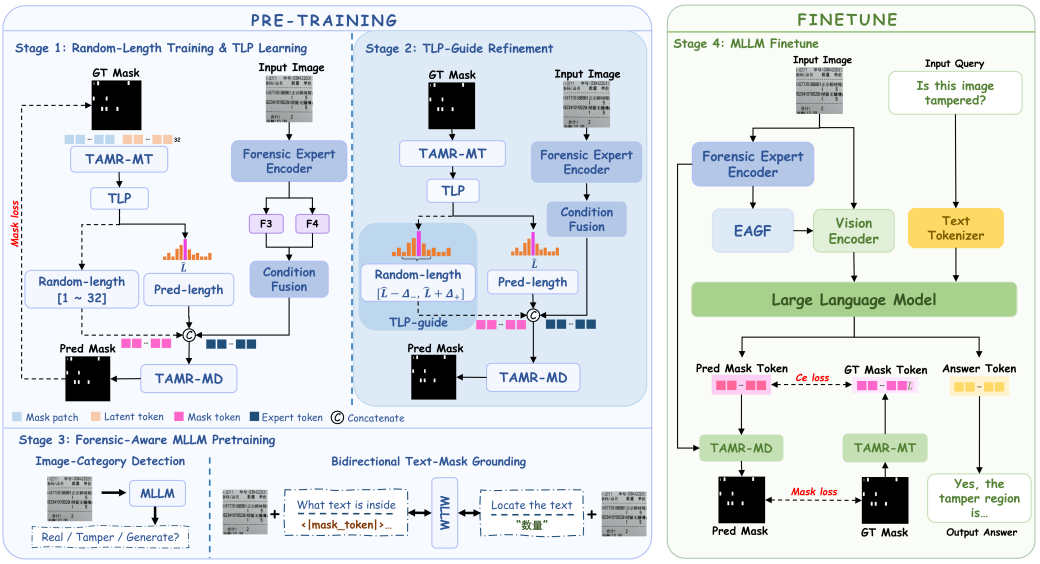


Figure 2: Overview of TraceTok's four-stage training pipeline.

stream through EAGF. The visual encoder, language model, and TAMR-MD are then fine-tuned for joint textual response generation and mask recovery.

3.2 Tamper-Aware Adaptive-Length Mask Reconstruction

LISA-style methods decode a mask from a special token through an external segmentation decoder, whereas Text-style methods return a textual bounding box; TraceTok serializes the target mask into discrete tokens and applies autoregressive token-label supervision. ALToLLM (Wang et al., 2025a) provides a mask-token reconstruction path. Given a target tampering mask M, its Mask Tokenizer (MT) encodes it into a quantized mask-token sequence Z. The Token Length Predictor (TLP) predicts a retained prefix length $\widehat{L}$, and the mask de-tokenizer (TAMR-MD) reconstructs the mask from the retained prefix conditioned on the feature map X_{SAM} produced by the SAM image encoder.

However, the SAM feature map emphasizes semantically salient objects rather than manipulation-specific regions, so the decoder may miss small tampered-text areas and reconstruct their irregular shapes inaccurately. TLP provides an initial length estimate but may still require finer allocation across masks with different spatial structures. We therefore propose Tamper-Aware Adaptive-Length Mask Reconstruction (TAMR), which introduces image-side forensic features for reconstruction and uses TLP-guided bounded sampling to refine token allocation.

To provide manipulation-related cues during reconstruction, TAMR conditions TAMR-MD on image-side forensic features. For an input image I, let $F_e = \mathcal{E}_f(I)$ denote the expert feature supplied to TAMR-MD by the frozen forensic expert encoder, and let $\widetilde{F}_e$ denote its spatially aligned and channel-projected form on the TAMR-MD query grid. The same encoder also produces stage-specific feature maps F_s; EAGF uses F_3 and F_4 from stages 3 and 4, respectively. For a retained prefix $Z_{1:L}$, the expert-conditioned query and the reconstructed mask $\widehat{M}$ are

$$Q_e = Q + \widetilde{F}_e, \qquad \widehat{M} = \mathcal{D}_m(\mathrm{Concat}(Q_e, Z_{1:L}))\,, \tag{1}$$

where $\mathcal{D}_m$ denotes the mask decoder implemented by TAMR-MD, Q is the base mask query used by TAMR-MD, $Z_{1:L}$ is the retained discrete mask-token prefix, and $\mathrm{Concat}(\cdot,\cdot)$ denotes concatenation along the token dimension.

To refine the retained length for tampered-text masks, TAMR uses TLP-guided bounded sampling. As shown in Fig. 2, in Stage 1, TLP is frozen while the reconstructor is trained with a prefix length sampled uniformly from 1 to 32 tokens. The reconstructor is then fixed, and TLP is optimized through the reconstruction objective to learn an initial correspondence between mask complexity

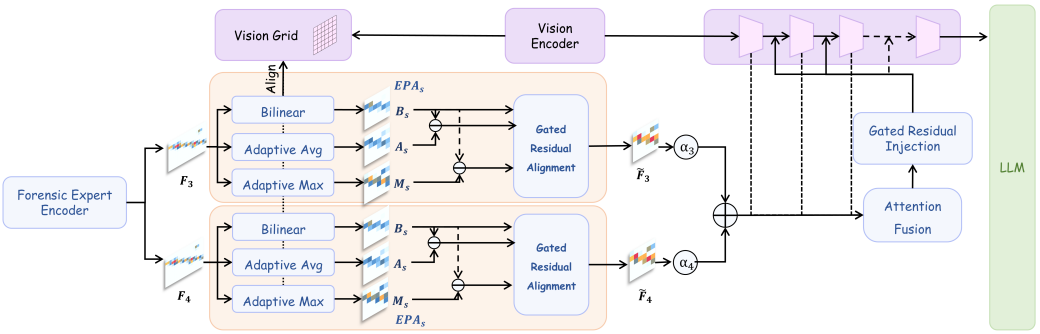


Figure 3: Overview of EAGF and its EPA-based alignment to the current visual-token grid.

and retained length. In Stage 2, the initialized TLP predicts $\widehat{L}$ for each mask. We define a bounded sampling window around this prediction and sample L within it:

$$L \sim \mathcal{U}_{\mathbb{Z}}\Big[\max(1, \widehat{L} - \Delta_{-}), \min(K, \widehat{L} + \Delta_{+})\Big], \tag{2}$$

where $K = 32$ is the maximum number of mask-token positions, Δ_{-} and Δ_{+} specify the lower and upper offsets of the sampling window, and $\mathcal{U}_{\mathbb{Z}}$ denotes the discrete uniform distribution. Stage 2 continues reconstruction training with lengths sampled from this TLP-guided window rather than from the full range. This rule preserves controlled randomness, adapts the retained prefix length to mask complexity, and encourages useful information to concentrate in earlier tokens.

Furthermore, although adaptive length selection and forensic conditioning improve reconstruction, the objective still weights all pixels equally, so a small tampered foreground can be dominated by background. We therefore use a foreground-weighted objective to reduce this dominance (Sec. S2).

3.3 Evidence-Preserving Adaptive Grid Fusion

General-purpose MLLMs provide useful semantic visual representations but are less sensitive to the low-level traces of text manipulation. A frozen forensic expert encoder supplies complementary forensic features, but its multistage feature maps vary with the input resolution, whereas the visual branch represents the image on a visual-token grid determined by the visual encoder. The two grids therefore rarely coincide, and a visual token and an expert response at the same index need not describe the same image region, so the two branches cannot be fused position by position without an alignment step.

To address this issue, we propose Evidence-Preserving Adaptive Grid Fusion (EAGF). As shown in Figure 3, the frozen expert encoder outputs the stage-3 and stage-4 feature maps F_3 and F_4. Each feature is sent to an Evidence-Preserving Aligner (EPA) and mapped to the current visual-token grid. The aligned features are then fused at corresponding spatial positions and injected into intermediate visual representations through a gated residual update. This process preserves the original visual representation while exposing the MLLM to complementary forensic evidence.

Specifically, we use a forensic expert trained on the corresponding dataset and freeze its encoder during MLLM training. F_3 and F_4 are the feature maps produced at expert stages 3 and 4, respectively. For the l-th injection location in the visual branch, let G_l denote the target visual-token grid. Interpolating F_s onto G_l is the most direct way to match the two grids, but it resamples neighbouring expert responses, which attenuates the localized traces that distinguish manipulated characters from their surroundings. Adaptive max pooling instead keeps only the strongest response in each grid cell, so weaker evidence inside the cell is discarded and the resulting map becomes spatially coarse. Tampered-text evidence can be concentrated in a small peak or spread across a region, so EPA takes the interpolated map as a geometric baseline and adds the cell-level average and peak responses as gated residuals. For each expert stage $s \in \{3, 4\}$, F_s is mapped to the target grid G_l by bilinear interpolation, adaptive average pooling, and adaptive max pooling, producing B_s, A_s, and M_s, respectively. EPA then computes

$$\bar{F}_s = B_s + \sigma(a_s)\phi_a(A_s - B_s) + \sigma(m_s)\phi_m(M_s - B_s), \tag{3}$$

where B_s is the bilinear-aligned baseline on G_l, A_s and M_s are the adaptive average- and max-response maps on the same target grid, and $\bar{F}_s$ denotes the aligned feature of expert stage s on G_l. The functions ϕ_a and ϕ_m project the two residual responses, while a_s and m_s are learnable gate parameters that control how much cell-level evidence each stage contributes. The residual branches are zero-initialized so that EPA initially preserves the bilinear alignment path.

After the feature maps from both stages have been aligned, EAGF selects their complementary information at each spatial position. Let x denote a spatial position on G_l, and let $V_l(x)$ be the current visual feature at that position. The fused evidence at that position and the stage weights are defined as

$$\begin{aligned} E_l(x) &= \sum_{s\in\{3,4\}} \alpha_s(x)\mathcal{V}_s(\bar{F}_s(x)), \\ \alpha_s(x) &= \mathrm{Softmax}_{s\in\{3,4\}}\left(\frac{\mathcal{Q}(V_l(x))^\top \mathcal{K}_s(\bar{F}_s(x))}{\sqrt{d}}\right), \end{aligned} \tag{4}$$

where $\mathcal{Q}$ maps the current visual feature to a query, while $\mathcal{K}_s$ and $\mathcal{V}_s$ map the aligned feature from expert stage s to a key and a value, respectively. The variable d denotes the key dimension. The weights $\alpha_s(x)$ are normalized only over the two expert stages at the same spatial position, so $\alpha_3(x)+\alpha_4(x) = 1$. The resulting $E_l(x)$ is the fused forensic evidence at that position. Finally, EAGF injects the fused evidence into the visual branch:

$$\widehat{V}_l = V_l + \gamma_l \Phi_l(\mathrm{Concat}(V_l, E_l)), \tag{5}$$

where Φ_l is the fusion projection, $\mathrm{Concat}(V_l, E_l)$ denotes channel-wise concatenation, γ_l is a lower-bounded learnable gate that scales the residual response and is clipped so that the injected evidence matches the magnitude of V_l. This operation updates the visual feature through a residual connection. EAGF does not generate a tampering mask; it provides richer forensic visual information for subsequent textual response generation and TAMR-MD mask reconstruction (Sec. S3).

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We evaluate TraceTok on TFR and RTM. For TFR, we report localization under two protocols (Qu et al., 2026): the Box Localization Protocol follows the TFR protocol of TextShield-R1, so that the results remain comparable with prior MLLM-based methods, while the Mask Localization Protocol uses the ground-truth tampering masks as localization targets and applies no OCR-based adjustment, because tampered text regions are usually not axis-aligned rectangles. TFR provides a training split and four evaluation subsets: an in-distribution Test split, a Cross-Image-Style (CIS) split with unseen image sources, a Cross-Tampering-Method (CTM) split with unseen tampering methods, and a Cross-Language (CL) split with unseen languages. Both protocols report IoU, precision, and recall on the Test, CIS, and CTM subsets and follow the TFR protocol, which averages the per-image values over each subset; additional CL results are reported in the supplementary material (Sec. S6). RTM follows its original text-manipulation split and evaluation protocol (Luo et al., 2025), where IoU, precision, recall, and F1 are computed from counts accumulated over the whole split. All metrics are reported in percentage form, and the best and second-best results in each table are shown in bold and underlined.

Implementation and training. TraceTok uses InternVL2.5-8B (Chen et al., 2024a) and a frozen SegFormer-MiT-B2 forensic expert (Xie et al., 2021). The MLLM is trained for 10 epochs with AdamW, learning rate 4×10^{-5}, weight decay 0.05, and batch size 2 per GPU. TAMR is trained in two stages with lengths sampled from $[\widehat{L} - 12, \widehat{L} + 4]$ and clipped to $[1, 32]$. Full pre-training, expert-training, and TAMR configurations are given in the supplementary material (Sec. S1).

Baselines. For TFR, the Box Localization Protocol compares with fine-tuned InternVL3 (Zhu et al., 2025a), Qwen2.5-VL (Bai et al., 2025), and TextShield-R1 (Qu et al., 2026), while the Mask Localization Protocol additionally includes FakeShield (Xu et al., 2025b), SIDA (Huang et al., 2025a), ALToLLM (Wang et al., 2025a), and ForensicsTok (Xu et al., 2026). For RTM, we compare with

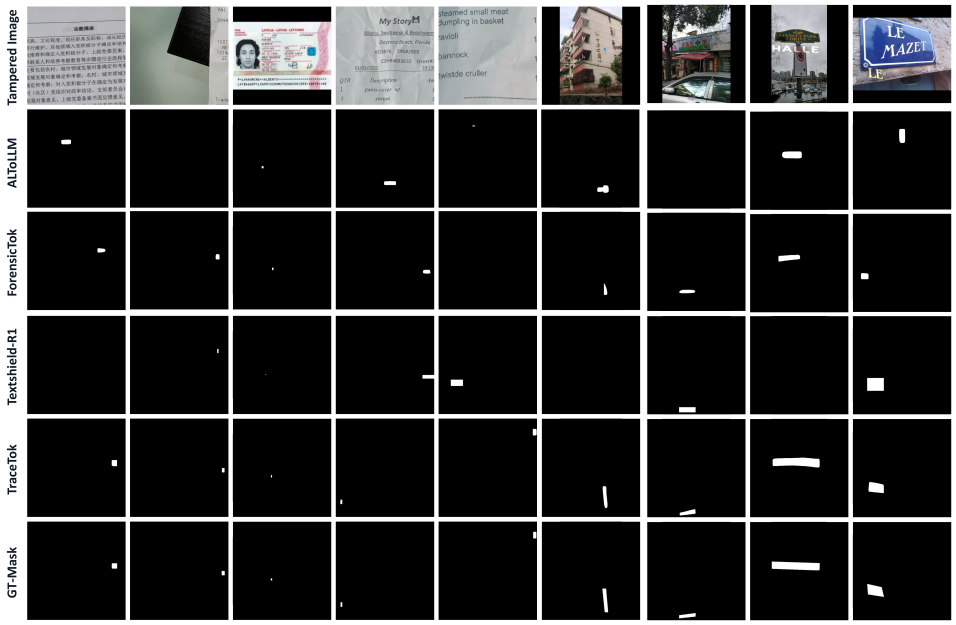


Figure 4: Qualitative comparison between TraceTok and competing methods.

the task-specific localizers reported in its original protocol and additionally with TVSIP (Xu et al., 2025a), an MLLM-based tampered-text localization method under the same protocol.

The off-the-shelf and fine-tuned results for GPT-4o, InternVL3, and Qwen2.5-VL, together with the FakeShield and SIDA results, are taken from the TFR evaluation of TextShield-R1; the full off-the-shelf comparison is given in the supplementary material (Sec. S5). Results marked with $*$ are obtained by evaluating the official pre-trained weights, and results marked with † are retrained by us under the corresponding protocol; unmarked baseline results retain their reported values.

4.2 Main Results on TFR

Table 1 reports TFR localization under the two protocols defined above. For TraceTok under the box protocol, we take the minimum enclosing box of the binarized predicted mask and refine it with OCR following the TextShield-R1 protocol, which is the protocol used for the compared methods (Sec. S5). Figure 4 shows how the methods differ on fine tampered-text regions. TextShield-R1* misses several edits or reduces them to compact rectangular patches, whereas ForensicsTok† locates the approximate edited area but its predicted masks often shift away from the target and omit parts of the text strokes. TraceTok follows the extent of thin and discontinuous regions more closely, particularly in the text-line and sign examples, although very small edits remain difficult to recover.

The visual difference reflects the models' localization mechanisms. A bounding-box prediction can locate a suspicious region but cannot specify its irregular pixel boundary. TraceTok instead generates mask tokens supervised by the target mask: TAMR reconstructs small foreground structures using forensic features and an adaptive retained length, while EAGF aligns expert features with the visual-token grid before injection. This preserves local evidence that box predictions lose at stroke boundaries.

Across the three TFR subsets, TraceTok reaches an average Mask IoU of 62.2, compared with 49.5 for TextShield-R1* and 35.9 for ForensicsTok†. The visual examples indicate where this gain arises: more complete recovery of thin and discontinuous tampered text rather than only a coarse indication of the affected region. The gain is not uniform. TraceTok improves substantially on large, irregular regions, where box outputs fail. On small regions, it mainly recovers thin or disconnected strokes missed by ALToLLM and remains competitive with task-specific models. A dense SegFormer expert remains ahead on Test and CIS, so the mask-token path approaches rather than replaces an expert trained with dense pixel supervision (Sec. S6).

Table 1: TFR localization results under the *Box Localization Protocol* and the *Mask Localization Protocol*. IoU, precision (P), and recall (R) are reported on the Test, CIS, and CTM subsets. Best and second-best results are shown in bold and underlined, respectively.

Box Localization Protocol									
Method	Test			CIS			CTM		
	Bbox IoU	P	R	Bbox IoU	P	R	Bbox IoU	P	R
Fine-tuned MLLM models									
InternVL3-2B	10.3	-	-	21.4	-	-	26.2	-	-
InternVL3-8B	15.4	-	-	25.2	-	-	31.4	-	-
Qwen2.5-VL-3B	11.6	-	-	20.9	-	-	25.6	-	-
Qwen2.5-VL-7B	18.2	-	-	26.5	-	-	34.2	-	-
TextShield-R1*	<u>50.8</u>	<u>53.1</u>	<u>53.6</u>	<u>56.4</u>	<u>58.1</u>	<u>64.5</u>	<u>66.4</u>	<u>68.0</u>	<u>69.5</u>
Our method									
TraceTok	**58.2**	**62.4**	**64.8**	**61.6**	**68.7**	**71.6**	**85.0**	**86.5**	**86.0**
Mask Localization Protocol									
Method	Test			CIS			CTM		
	Mask IoU	P	R	Mask IoU	P	R	Mask IoU	P	R
Fine-tuned mask baselines									
FakeShield	7.6	-	-	15.0	-	-	21.8	-	-
SIDA	7.7	-	-	15.1	-	-	21.8	-	-
MLLM and mask-token baselines									
TextShield-R1*	<u>39.0</u>	43.7	<u>42.8</u>	<u>52.5</u>	<u>57.1</u>	<u>59.5</u>	<u>57.0</u>	<u>60.9</u>	<u>66.2</u>
ALToLLM†	18.9	28.4	25.9	15.2	25.2	22.2	27.3	38.7	42.5
ForensicsTok†	37.0	<u>43.9</u>	42.2	29.9	39.2	35.7	40.7	60.4	46.3
Our method									
TraceTok	**53.4**	**60.1**	**60.1**	**60.2**	**68.6**	**70.8**	**73.1**	**79.8**	**79.8**

* Evaluated with the official pre-trained weights. † Retrained by us under the corresponding protocol. A dash means P or R was not reported in the source results reused from TextShield-R1.

4.3 TAMR Reconstruction Quality on TFR

Table 2 evaluates mask reconstruction independently of localization on the TFR Test, CIS, and CTM subsets and on RTM. Each method uses its tokenizer–reconstructor pair and reports retained tokens (#Tok.), Mask IoU, and F1. The baselines are HiMTok (Wang et al., 2025b) and ALToLLM (Wang et al., 2025a): ALToLLM is evaluated by re-inferring its released checkpoint, while ALToLLM† is retrained by us on the corresponding subsets.

TraceTok reaches the best reconstruction quality with the shortest sequences on all subsets. Against ALToLLM†, it improves Mask IoU/F1 by +3.6/+3.2 on Test, +2.3/+2.0 on CIS, +0.7/+0.6 on CTM, and +3.8/+3.4 on RTM, while using 0.9–1.2 fewer tokens. The gain therefore comes from reconstruction rather than a larger token budget.

The gains are largest on Test and RTM, which contain many small or multiple tampered regions. CTM is dominated by larger, contiguous masks, so the margin there is small. TAMR's forensic conditioning and TLP-guided budget allocation are what preserve local evidence there.

Table 2: Mask reconstruction comparison against autoregressive mask-token methods on TFR and RTM.

Method	Test			CIS			CTM			RTM		
	#Tok. ↓	IoU ↑	F1 ↑	#Tok. ↓	IoU ↑	F1 ↑	#Tok. ↓	IoU ↑	F1 ↑	#Tok. ↓	IoU ↑	F1 ↑
HiMTok	32.0	76.7	82.3	32.0	84.8	90.1	32.0	91.4	94.3	32.0	70.4	78.7
ALToLLM	17.4	77.8	84.4	17.6	83.0	89.1	17.6	95.7	97.7	19.6	75.3	83.3
ALToLLM†	<u>13.8</u>	<u>86.6</u>	<u>90.6</u>	<u>14.1</u>	<u>91.3</u>	<u>94.2</u>	<u>15.9</u>	<u>97.2</u>	<u>98.3</u>	<u>18.6</u>	<u>81.6</u>	<u>86.9</u>
TraceTok	**12.9**	**90.2**	**93.8**	**13.2**	**93.6**	**96.2**	**14.8**	**97.9**	**98.9**	**17.4**	**85.4**	**90.3**

† Retrained by us on the corresponding subsets; the unmarked ALToLLM row is obtained by re-inferring its released checkpoint. Scores in this table are computed per image and averaged over each subset.

4.4 Main Results on RTM

To evaluate localization on realistic text manipulation, Table 3 compares TraceTok with task-specific forensic models and MLLM-based methods under the RTM protocol. The compared models span generic segmentation backbones (Liu et al., 2021; 2022b), image-manipulation detectors (Dong

et al., 2023; Zhu et al., 2025b), and tampered-text localizers (Luo et al., 2025). TVSIP and TraceTok are reported under the RTM protocol; TraceTok is comparable to the strongest baselines: 20.9 IoU and 34.6 F1 against 19.7/32.9 for the task-specific ASC-Former and 20.6/34.2 for the MLLM-based TVSIP; its precision is lower (42.5 vs. 45.1) but its recall is higher (29.2 vs. 27.5).

The advantage over TVSIP is architectural rather than numerical. TVSIP decouples the pipeline: an MLLM predicts bounding boxes, OCR rectifies them, and the rectified boxes are fused with expert masks, so its localization inherits their errors. TraceTok instead generates both the explanatory text and the mask-token sequence in the same autoregressive pass, and reconstructs the mask with TAMR; the mask is supervised directly by the target mask and needs no box or OCR proposal. One possible explanation for the higher recall is that the mask-token path covers small or weakly contrasted regions that a box-plus-OCR pipeline drops, at the cost of a slightly lower precision. TraceTok therefore matches the strongest task-specific localizer while avoiding box and OCR proposals.

Table 3: Localization comparison against state-of-the-art methods under the RTM protocol.

Method	Venue	IoU	P	R	F1
Swin-UPer (Liu et al., 2021)	ICCV'21	17.0	41.8	22.3	29.1
ConvNeXt (Liu et al., 2022b)	CVPR'22	16.2	35.8	22.8	27.9
SegFormer (Xie et al., 2021)	NeurIPS'21	19.4	**56.5**	22.8	32.5
Mask2Former (Cheng et al., 2022)	CVPR'22	12.4	19.1	25.9	22.0
PSCC-Net (Liu et al., 2022a)	TCSVT'22	3.3	3.6	**30.3**	6.4
CAT-Net V2 (Kwon et al., 2022)	IJCV'22	13.7	25.8	22.6	24.1
MVSS-Net++ (Dong et al., 2023)	TPAMI'23	4.3	17.1	5.4	8.2
Mesorch (Zhu et al., 2025b)	AAAI'25	12.1	19.9	23.6	21.6
DTD (Qu et al., 2023)	CVPR'23	6.5	11.9	12.5	12.2
FFDN (Chen et al., 2024b)	ECCV'24	10.9	16.9	23.5	19.7
ASC-Former (Luo et al., 2025)	PR'25	19.7	50.4	24.4	32.9
TVSIP* (Xu et al., 2025a)	ACM MM'25	20.6	45.1	27.5	34.2
TraceTok	This work	**20.9**	42.5	29.2	**34.6**

* Evaluated with the official pre-trained weights.

4.5 Component Ablations

To evaluate how each component contributes to TraceTok, Table 4 ablates TAMR, MLLM pre-training (MLLM PT), and EAGF on the TFR subsets. Row 1 is the baseline, row 2 enables TAMR, and rows 3–5 test the remaining combinations; jointly enabling MLLM PT and EAGF (row 1 vs. row 3) gives the largest improvement. TAMR and MLLM PT then provide complementary gains: TAMR helps most on CIS and Test, where tampered regions are small and thin, whereas MLLM PT helps most on CIS and CTM, whose masks depend more on the text–mask correspondence. A per-subset analysis is given in the supplementary material (Sec. S6). With TAMR enabled and MLLM pre-training disabled, adding EAGF (row 2 vs. row 4) improves Mask IoU by 26.6, 22.2, and 25.1 points on Test, CIS, and CTM, respectively.

Table 4: Ablation study of TAMR, MLLM pre-training (MLLM PT), and EAGF on the TFR subsets.

Num.	TAMR	MLLM PT	EAGF	Test		CIS		CTM	
				F1	IoU	F1	IoU	F1	IoU
1	–	–	–	22.3	18.9	21.6	15.2	37.9	27.3
2	✓	–	–	25.7	22.7	23.5	19.8	38.6	27.9
3	–	✓	✓	54.6	47.6	62.2	53.3	75.5	69.5
4	✓	–	✓	55.7	49.3	50.4	42.0	61.2	53.0
5	✓	✓	✓	**59.3**	**53.4**	**68.0**	**60.2**	**78.8**	**73.1**

5 Conclusion

TraceTok formulates tampered-text localization as adaptive-length mask-token generation. TAMR reconstructs small, fragmented, or disconnected tampering masks from compact token sequences, while EAGF injects aligned multistage expert features into the visual stream. A progressive training recipe lets one autoregressive path generate the textual explanation and a pixel-level mask.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, software implementation and experiment-support code, LaTeX and bibliography formatting, and the iterative development of a conceptual figure. They were also used to provide feedback on the organization and presentation of the work. The authors reviewed all AI-assisted text, code, formatting, and visual material against the experiment records and cited literature. Generative AI was not used to generate experimental data, compute or modify the reported quantitative results, or make the final interpretation of results. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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

Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Jun Lan, Xing Zheng, Yaohui Li, Changhua Meng, Huijia Zhu, and Weiqiang Wang. DiffUTE: Universal text editing diffusion model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 63062–63074, 2023.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Hanming Lv, Yi Wang, Wenqi Shao, Peiyu Chu, Yuchao Tu, Tong He, Zhiyong Wu, Wenhai Deng, Kaipeng Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Yuang Zhang, Wen Gao, Hongsheng Li, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl2.5: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhongxi Chen, Shen Chen, Taiping Yao, Ke Sun, Shouhong Ding, Xianming Lin, Liujuan Cao, and Rongrong Ji. Enhancing tampered text detection through frequency feature fusion and decomposition. In *European Conference on Computer Vision*, pp. 200–217, 2024b.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1280–1289, 2022.

Chenbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2023.

Li Dong, Weipeng Liang, and Rangding Wang. Robust text image tampering localization via forgery traces enhancement and multiscale attention. *IEEE Transactions on Consumer Electronics*, 70(1): 3495–3507, 2024.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/cvpr52729.2023.01974.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Zhenglin Huang, Xiangtai Li, Xi Yang, Bei Peng, Xiaowei Huang, Baoyuan Wu, Dacheng Tao, Ming-Hsuan Yang, and Guangliang Cheng. So-Fake: Benchmarking and explaining social media image forgery detection. *arXiv preprint arXiv:2505.18660*, 2025b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12985–12996, 2024.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. In *International Conference on Learning Representations*, 2025.

Songze Li, Yunfei Guo, Shen Chen, Bin Li, Kaiqing Lin, Changsheng Chen, Haodong Li, Taiping Yao, and Shouhong Ding. DITL^2: Dual-stage invariance transfer learning for generalizable document image tampering localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 82–91, 2025a. doi: 10.1145/3746027.3754857.

Weixiang Li, Bin Li, Kengtao Zheng, Songze Li, and Haoliang Li. Document image forgery detection and localization in desensitization scenarios. *Signal Processing*, 238:110123, 2026. doi: 10.1016/j.sigpro.2025.110123.

Youjie Li, Shiqiang Zheng, Guijia Zhang, Qifeng Chen, and Changsheng Chen. DocAI-TL: Structured document tampering localization with DocAI model. In *International Conference on Document Analysis and Recognition*, pp. 311–328, 2025b. doi: 10.1007/978-3-032-04627-7_18.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9992–10002, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11966–11976, 2022b.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157:110828, 2025. doi: 10.1016/j.patcog.2024.110828.

Tung Nguyen, Yuxin Wen, Preston K. Robinette, Taylor T. Johnson, Tom Goldstein, Anh Tran, and Khoi Nguyen. EditScout: Locating forged regions from diffusion-based edited images with multimodal LLM. *arXiv preprint arXiv:2412.03809*, 2024.

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

Chenfan Qu, Jian Liu, Haoxing Chen, Baihan Yu, Jingjing Liu, Weiqiang Wang, and Lianwen Jin. Textsleuth: Towards explainable tampered text detection. *arXiv preprint arXiv:2412.14816*, 2024.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 694–702, 2025. doi: 10.1609/aaai.v39i1.32051.

Chenfan Qu, Yiwu Zhong, Jian Liu, Xuekang Zhu, Bohan Yu, and Lianwen Jin. Textshield-r1: Reinforced reasoning for tampered text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8621–8629, 2026. doi: 10.1609/aaai.v40i10.37814.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26374–26383, 2024.

Huiru Shao, Zhuang Qian, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. Delving into adversarial robustness on document tampering localization. In *European Conference on Computer Vision*, pp. 290–306, 2024. doi: 10.1007/978-3-031-73650-6_17.

Yalin Song, Wenbin Jiang, Xiuli Chai, Zhihua Gan, Mengyuan Zhou, and Lei Chen. Cross-attention based two-branch networks for document image forgery localization in the metaverse. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(2):1–24, 2025. doi: 10.1145/3686158.

Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? SparseViT: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7024–7032, 2025. doi: 10.1609/aaai.v39i7.32754.

Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. ForgerySleuth: Empowering multimodal large language models for image manipulation detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 169963–170002, 2025. doi: 10.52202/085713-5121.

Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. AnyText2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024.

Lingfeng Wang, Hualing Lin, Senda Chen, Tao Wang, Changxu Cheng, Yangyang Zhong, Dong Zheng, and Wuyue Zhao. Alto: Adaptive-length tokenizer for autoregressive mask generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 18090–18114, 2025a. doi: 10.52202/085713-0542.

Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. HiMTok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23267–23278, 2025b. doi: 10.1109/iccv51701.2025.02160.

Yuxin Wang, Hongtao Xie, Mengting Xing, Jing Wang, Shenggao Zhu, and Yongdong Zhang. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, pp. 215–232, 2022.

Kahim Wong, Jicheng Zhou, Haiwei Wu, Yain-Whar Si, and Jiantao Zhou. ADCD-Net: Robust document image forgery localization via adaptive DCT feature and hierarchical content disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19280–19289, 2025. doi: 10.1109/iccv51701.2025.01792.

Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the ACM International Conference on Multimedia*, pp. 1500–1508, 2019.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3858–3869, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Guitao Xu, Ziqi Yi, Peirong Zhang, Jiahuan Cao, Shihang Wu, and Lianwen Jin. From pixels to semantics: A novel mllm-driven approach for explainable tampered text detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 757–766, 2025a. doi: 10.1145/3746027.3755803.

Lei Xu, Haowei Wang, Shen Chen, Taiping Yao, Bin Li, and Changsheng Chen. Forensic-stok: Forensics-guided tokenized modeling for image tampering localization. *arXiv preprint arXiv:2606.24538*, 2026.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, 2025b.

Ziqi Yi, Guitao Xu, Shihang Wu, Peirong Zhang, and Lianwen Jin. Frequency mining empowered by text aggregation: A new perspective on document image tampering detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1471–1479, 2026. doi: 10.1609/aaai.v40i2.37122.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4096.

Fanrui Zhang, Jiawei Liu, Jiaying Zhu, Esther Sun, Dong Li, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: A multimodal llm for interpretable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024a.

Keyang Zhang, Chenqi Kong, Hui Liu, Bo Ding, Xinghao Jiang, and Haoliang Li. Propose and rectify: A forensics-driven MLLM framework for image manipulation localization. *IEEE Transactions on Information Forensics and Security*, 21:3396–3409, 2026.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91, 2024b.

Shiqiang Zheng, Changsheng Chen, Shen Chen, Taiping Yao, Shouhong Ding, Bin Li, and Jiwu Huang. Generalized document tampering localization via color and semantic disentanglement. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(1):1279–1292, 2026. doi: 10.1109/tcsvt.2025.3597602.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025. doi: 10.1109/iccv51701.2025.01742.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Ji, Hao Li, Linglin Jing, Siyuan Liu, Xiangyu Zhao, Lu Chen, Kaipeng Zhang, Wenhai Wang, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale and hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11022–11030, 2025b. doi: 10.1609/aaai.v39i10.33198.

APPENDIX: LIMITATIONS

TraceTok localizes tampered text but does not identify the manipulation type or the semantic intent behind it, and it depends on a frozen forensic expert: when the expert misses a manipulation, the mask-token path inherits that error. Keeping a second encoder also makes inference costlier than a single-MLLM pipeline. Our evaluation covers localization on TFR and RTM; the classification task and the quality of the generated textual response are not evaluated, and all reported numbers come from a single training run per setting.