# TextSleuth: Towards Explainable Tampered Text Detection

**Anonymous ACL submission**


## Abstract

Recent advancements in tampered text detection has attracted increasing attention due to its essential role in information security. Although existing methods can detect the tampered text region, the detection lacks convincing interpretation and clarity, making the prediction unreliable. To address this problem, we propose to explain the basis of tampered text detection with natural language via large multimodal models. To bridge the data gap, we propose a large-scale, comprehensive dataset, ETTD, which contains both pixel-level annotations for tampered text region and natural language annotations describing the anomaly of the tampered text. Multiple novel methods are employed to improve the quality of our dataset. To further improve explainable tampered text detection, we propose a simple yet effective model called **TextSleuth**, which can detect tampered text with **both visual and semantic clues**, and shows **strong generalization across unfamiliar image styles and languages**. Extensive experiments on both the ETTD dataset and the public dataset have verified the effectiveness of the proposed methods. Our dataset and code will be made publicly available.


## 1 Introduction

With the rapid development of image processing technologies, sensitive text information can be more easily manipulated for malicious purposes, such as fraud, posing serious risks to information security (Dong et al., 2024). Consequently, tampered text detection has become a major research topic in recent years (Qu et al., 2024b). Existing works model tampered text detection as semantic segmentation (Shao et al., 2023) or object detection (Qu et al., 2024a). Despite the progress made in recent years, such fine-grained predictions are still black-box and lead to unreliable results.

To provide more reliable predictions for tampered text detection, we propose to leverage multimodal large models to both detect tampered text regions and explain the basis for their detection in natural language. Given the absence of dataset for interpretable tampered text detection, we construct the Explainable Tampered Text Detection (ETTD) dataset. To ensure the comprehensiveness of the data, we collect multilingual card images, document images and scene text images from the Internet and the existing text-rich datasets. We then perform text tampering on the collected data with various methods, including traditional methods copy-move, splicing, and the deep generative method DiffUTE (Chen et al., 2024a). Finally, we create 13,500 tampered text images with accurate pixel-level annotations of the tampered region and 11,500 authentic text images. The large-scale of our data notably alleviates the data hunger of deep models. The images are divided into ETTD-Train, ETTD-Test, ETTD-CD, and ETTD-ML. The ETTD-Test is the common test set. The ETTD-CD is a cross-domain test set with different distribution than the ETTD-Train. The ETTD-ML con-

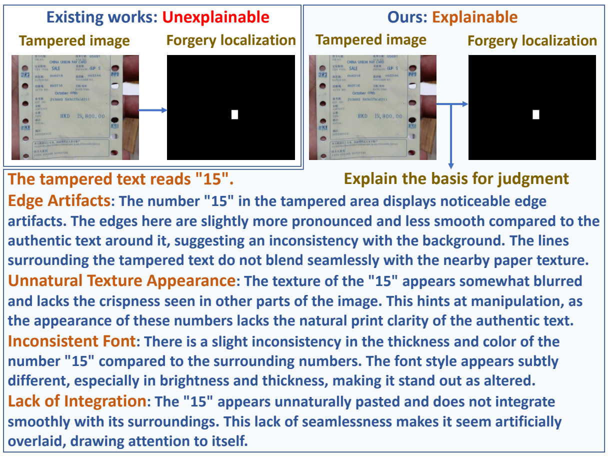



Figure 1: We propose to both detect the tampered text region and explain the basis for the detection in natural language, making the prediction more reliable. We construct the first dataset and propose a novel model for the explainable tampered text detection task.

tains tampered text in Arabic, Bengali, Japanese, and Korean. This differs from the ETTD-Train, which contains text in English and Chinese. The ETTD-CD and ETTD-ML enable the evaluation of cross-domain the multilingual performance of models, which is crucial in real world application.

With the obtained tampered text, we utilize GPT-4o to generate the description of both visual and linguistic anomalies caused by text tampering, and to generate the text recognition result for the target tampered text. To achieve this, we prompt the GPT-4o with a novel elaborate query, the tampered image and its corresponding mask annotation indicating the tampered region. However, since text is mostly dense and has similar location and shape, directly inputting the binary mask, as existing work (Xu et al., 2024) does will cause severe confusion to the GPT-4o, making it unclear which is the actual tampered text. To solve this problem, we propose to fuse the binary mask into the original tampered image with pixel-wise weighting. With the proposed fused mask prompt, the GPT-4o has a much better understanding of the location of the target region, which in turn significantly reduces the errors and obviously improves the annotation quality. In addition, the GPT-4o's output is not always correct and manual verification is costly. Inspired by the fact that incorrect detection of manipulated text leads to unclear perception and poor anomaly description, we further propose to address this issue by automatically filtering the annotation based on the OCR accuracy of the tampered text.

The tiny area and visual consistency of tampered text (Wang et al., 2022) pose multiple challenges for explainable tampered text detection, making it difficult for existing methods to achieve good enough performance. For example, misidentification of tampered text leads to incorrect anomaly description, difficulty in finding tampered text weakens the analysis quality, and increases the risk of overfitting to unrelated background styles. To this end, we propose a novel **simple-yet-effective** model termed as TextSleuth. Specifically, an extra RCNN (Ren et al., 2015) based text detection module initially scans the image and predicts the location of the tampered text with cascaded RoI heads. The initial prediction of tampered region is converted into a grounding prompt and fed into the large language model along with the image tokens and the original question to obtain the final prediction. The proposed two-stage analysis paradigm and auxiliary prompt in TextSleuth effectively minimizes errors, improves explanation quality and cross-domain generalization by drawing the model's special attention to the anomaly region and helping it to learn more general features. In addition, since the reference grounding comprehension task is mostly involved in the pre-training stage of large models (Chen et al., 2024b), the proposed auxiliary grounding prompt can reduce comprehension difficulty and alleviate forgetting.

Both our proposed ETTD dataset and TextSleuth model are the first efforts in the field of interpretable tampered text detection. Extensive experiments have confirmed that the proposed TextSleuth significantly improves upon the baseline model, outperforming existing methods by a large margin on both the proposed ETTD dataset and the public Tampered IC-13 (Wang et al., 2022) dataset, demonstrating strong in-domain and cross-domain generalization capabilities. In-depth analysis is also provided to inspire further work in the field of interpretable tampered text detection.

In summary, our main contribution is fourfold:

- We propose a novel **task**, explainable tampered text detection, which aims to provide reliable prediction by describing the anomalies of tampered text in natural language, serving as a pioneering effort in this field.
- We obtain the data annotation for this task by prompting GPT-4o with elaborate queries. We propose effective methods to improve the quality of the annotations. Based on these, we construct the ETTD dataset, which is **the first**, large-scale and comprehensive **dataset** for explainable tampered text detection.
- We propose **a pioneering** multimodal large **model** TextSleuth for interpretable tampered text detection, which achieves state-of-the-art performance with a two-stage analysis paradigm and a novel auxiliary prompt.
- Extensive experiments are conducted. Valuable inspiring conclusions and insights are provided through in-depth analysis.

## 2 Related works

### 2.1 Tampered Text Detection

Early work on tampered text detection is achieved by printer classification (Lampert et al., 2006) or template matching (Ahmed and Shafait, 2014), which is only applicable to scanned documents and does not work well for photographed documents (Dong et al., 2024). DTD (Qu et al., 2023) is proposed to detect visually consistent tampering in documents through examining the continuity of the block artifacts grids. CAFTB-Net (Song et al., 2024) benefits from noise-domain modeling and cross-attention mechanism. DTL (Shao et al., 2025) improves model robustness with latent manifold adversarial training. Despite the progress made in recent years, existing work on tampered text detection can still only localize the tampered region in an unreliable black-box manner, unable to explain the judgement basis in natural language.

### 2.2 Explainable Image Forgery Detection

Recently, several works achieve explainable image forensics through multimodal large language models. FFAA (Huang et al., 2024) utilizes GPT-4o to generate detailed basis description DeepFake artifacts. FakeShield (Xu et al., 2024) leverages GPT-4o to create anomaly description for natural style image forgery. ForgeryGPT (Li et al., 2024) improves interpretable natural image forensics with binary mask prompt. ForgerySleuth (Sun et al., 2024) obtains hierarchical forgery description annotation with the proposed Chain-of-Clues. Despite the progress made, none of the existing work achieves interpretable forensics on tampered text detection. Due to the tiny size and visual consistency of tampered text (Qu et al., 2023), natural image forgery detection methods struggle with tampered text detection (Luo et al., 2024). It is crucial to develop explainable tampered text detection techniques for reliable text image forensics.

## 3 ETTD Dataset

To fill in the data gap for explainable tampered text detection dataset, we construct a large-scale comprehensive dataset called ETTD.

### 3.1 Text Tampering

To ensure the comprehensiveness of our dataset, we collect multilingual document and card images from the Internet and scene text images from the existing datasets (listed in Table 1). We then forge some of the collected images with the widely-used methods, copy-move and splicing. Poisson Blending (Pérez et al., 2023) is employed to reduce visual inconsistency. To further improve the data diversity, we manually edit the text with DiffUTE (Chen et al., 2024a), a latest diffusion model for realistic tampered text generation.

### 3.2 Anomaly Description Generation

As shown in Figure 2, we leverage the GPT-4o to generate the description of both visual and linguistic anomalies caused by text tampering. Given the different features between tampered text and tampered natural objects (Wang et al., 2022), the textual queries in existing works (Xu et al., 2024) can not work well for tampered text (e.g. "unnatural depth" is usually observed in tampered natural objects but not in tampered text). To address this issue, we propose an elaborate query that inspires the GPT-4o to analyze anomalies for tampered text on six major perspectives, covering texture, integration, alignment, edge artifacts, text coherence, font, as shown in Figure 2. The detailed query is presented in the Appendix.

We then input this elaborate query along with the tampered image and its corresponding mask annotation into the GPT-4o. However, due to the similarity in location and shape of the text instances in an image, directly inputting the binary mask as done in existing work (Xu et al., 2024) will cause considerable confusion to the GPT-4o. As shown in Fig. 3, the annotator model usually struggles to identify the target text with the binary mask, often mis-detecting a nearby authentic text as a fake text. Analyzing anomaly on authentic text undoubtedly produces incorrect anomaly descriptions. To address this issue, we propose the fused mask prompt, where the original image is fused with the binary mask by pixel-wise weighting. Specifically, given the input image $I \in \mathbb{R}^{H,W,3}$ and the binary mask annotation $M \in \mathbb{B}^{H,W}$, $\mathbb{B} \in \{0,1\}$, the fused mask prompt $M^{fused} \in \mathbb{R}^{H,W,3}$ can be formulated as $M^{fused} = I * \lambda_1 + M * \lambda_2$. We set $\lambda_1$ and $\lambda_2$ to 0.5 in practical. With the proposed fused mask prompt, the annotator can clearly recognize the tampered text on the target region and better understand where the target region is by referring to the surrounding content. The proposed method significantly reduces hallucination and errors caused by frequent confusion.

Since the responses of GPT-4o are not always correct, directly using the GPT-4o responses as annotations leads to poor data quality, while manually verifying the annotation is costly. To this end, we propose an automatic filtering method to discard unsatisfactory responses. We empirically find that, the anomaly description from the GPT-4o is also mostly accurate when the GPT-4o can correctly recognize the tampered text. This means that the

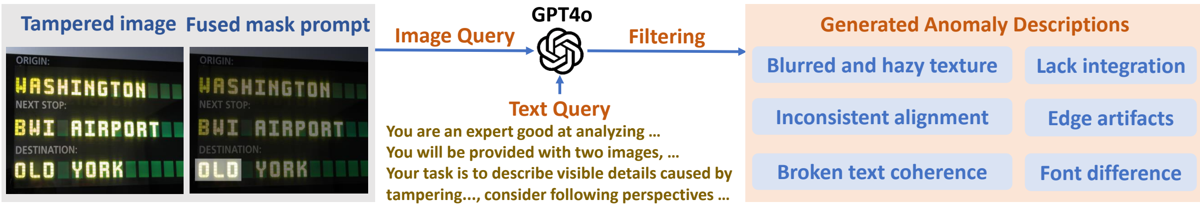



Figure 2: The pipeline for obtaining the textual anomaly description for the tampered text.

| Dataset | Image types | Image source | Languages | Tampering Types (# of samples) | Real num. | Forged Area |
|---|---|---|---|---|---|---|
| ETTD-Train | Documents, ID cards, scene texts, etc. | Internet, ICDAR2013, ICDAR2017, LSVT | EN, CH | Total (10400): DiffUTE (800), Copy-move (4800), Splicing (4800) | 9600 | 0.0268 |
| ETTD-Test | | | EN, CH | Total (600): DiffUTE (200), Copy-move (200), Splicing (200) | 400 | 0.0202 |
| ETTD-CD | scene text | ICDAR2013 | EN | Total (1000): Copy-move (500), Splicing (500) | 500 | 0.0608 |
| ETTD-ML | scene text | ICDAR2017 | ARA, BEN, JAN, KOR | Total (1500): Copy-move (500), Splicing (1000) | 1000 | 0.0415 |

Table 1: A brief summary of the ETTD dataset statistics. 'EN' denotes English, 'CH' denotes Chinese, 'ARA' denotes Arabic, 'BEN' denotes Bengali, 'JAN' denotes Japanese and 'KOR' denotes Korean. 'Real num.' denotes the number of authentic images. 'Forged Area' denotes the area ratio of tampered text.

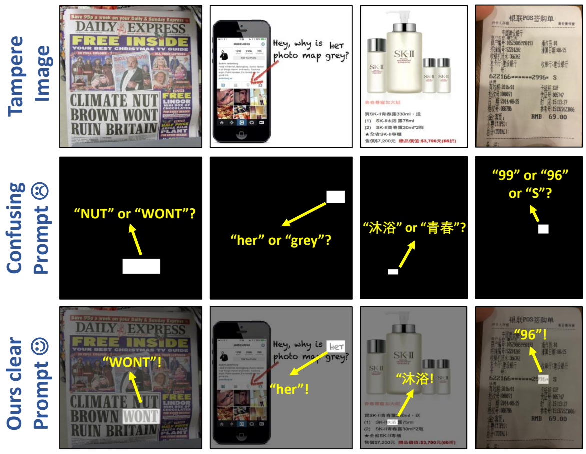



Figure 3: The binary mask prompt as in existing work is confusing in text images. In contrast, our proposed fused mask prompt clearly indicates the content and the exact location of the tampered text.

GPT-4o is clear about the location of the tampered text and the visual details of it. Based on this observation, we propose to automatically filter out the bad responses with tampered text OCR accuracy (Zhang et al., 2019) lower than 0.8. The OCR ground-truth is obtained from dataset annotation or OCR engine, and is used to replace the GPT-4o OCR in the remaining samples to ensure accuracy. The proposed method effectively improves the quality of anomaly description for tampered text in an automatic manner. For authentic text images, the textual description is set to "There is no tampered text in this image".

### 3.3 Dataset Summary

As shown in Tab. 1, there are 6,000 text images tampered by copy-move, 6,500 text images tampered by splicing and 1,000 text images tampered by DiffUTE in our ETTD dataset. The large-scale and comprehensiveness of our dataset can effectively alleviate the data hunger for deep forensic models. Another 11,500 images without text tampering serve as the authentic part. 20,000 images from the ETTD dataset are split as the training set (ETTD-Train), 1,000 images from the ETTD dataset are split as the test set (ETTD-Test) and another 1,500 images from the ETTD dataset are split as the cross-domain test set (ETTD-CD). The ETTD-CD consists of copy-move forgeries, splicing forgeries and authentic images from ICDAR2013, which are not included in ETTD-Train. Therefore, the ETTD-CD has a different data distribution from ETTD-Train and can evaluate model performance on unknown scenarios. 2500 text images with different languages from ETTD-Train are split as the multi-lingual test set ETTD-ML, which can evaluate model's multi-lingual performance on unseen languages. Accurate pixel-level annotations for tampered regions are provided to facilitate fine-grained analysis of the tampered text regions.

### 3.4 Dataset Highlights

The main highlights of our ETTD dataset include:

**Diverse.** As shown in Fig. 3 and Tab. 1, our dataset contains text with diverse fonts, lighting, backgrounds, and so on. The texts in our dataset include handwritten, printed, and scene text. They are tampered using both traditional and AIGC methods.

**Comprehensive.** Our ETTD-CD can evaluate model performance across unfamiliar image styles and our ETTD-ML can evaluate model performance across multiple unfamiliar languages. Our dataset ensures a comprehensive and thorough evaluation of explainable tampered text detection.

**Large Scale.** Our dataset contains 13,500 tampered text images and 11, 500 real text images, providing sufficient training and test samples.

## 4 TextSleuth

The tampered text is mostly small in size and the visual anomalies are often subtle (Luo et al., 2024). Consequently, two major challenges are emerged for interpretable tampered text detection: 1. The multimodal large models suffer from misidentifying the tampered text, resulting in incorrect anomaly description. 2. The models are more likely to be disturbed by the irrelevant background style, which weakens their generalization on unseen tampering methods and scenarios. To this end, we propose a **simple-yet-effective** model termed as TextSleuth, which overcomes the above challenges through a two-stage analysis paradigm and a novel reference grounding auxiliary prompt.

As shown in Figure 4, given an input image, the suspected tampered text region is initially detected by a Swin-Transformer based detector with cascaded RoI heads (Cai and Vasconcelos, 2018). The predicted coordinates are then normalized to 0-1000 and are converted to the reference grounding auxiliary prompt "The suspected tampered text $\langle box\rangle[[x_{min}, y_{min}, x_{max}, y_{max}]]\langle/box\rangle$". Given that the reference grounding comprehension task is involved in the pre-training stage of most large models (Wang et al., 2024), the large language model can effortlessly comprehend the target location in the proposed auxiliary prompt. In the auxiliary prompt, the large language model naturally pays special attention to the region represented by the coordinates, as it has learned in its pre-training stage. This differs from existing work (Li et al., 2024) that forces the model to look at the suspected region with binary mask embeddings, which is confusing in indicating tampered text, violates the pre-training paradigm and causes more forgetting. The auxiliary prompt is fed into the large language model along with the image tokens and the original question, to obtain the recognition and describe the anomaly for tampered text.

Despite its simplicity, the proposed method effectively addresses the major challenges in explainable tampered text detection: 1. The initial prediction of the suspect region significantly reduces the misidentification of the tampered region and reduces hallucination. 2. By focusing on the tampered region, the model gets rid of the interference from unrelated background styles, learns more general features during training, and thereby perform better on unseen tampering methods and scenarios.

## 5 Experiments

We conduct experiments on both the proposed ETTD dataset and the public Tampered IC-13 dataset (Wang et al., 2022) with multiple advanced multimodal LLMs, including GPT-4o (OpenAI, 2024), Yi-VL-6B (AI et al., 2024), DeepSeekVL-7B (Lu et al., 2024), MiniCPM-V2.5 (Hu et al., 2024), the 1B to 8B versions of Intern2-VL (Chen et al., 2024b) and the 2B, 7B versions of Qwen2-VL (Wang et al., 2024). We fine-tune all models except GPT-4o on the ETTD-Train for 5 epochs with the same settings.

### 5.1 Evaluation Metric

To evaluate the similarity between the predicted anomaly description and the textual annotation, we calculate the OCR accuracy (Zhang et al., 2019) for tampered text recognition and the word-to-vector-based (Mikolov et al., 2018) cosine similarity, Rouge-L, BELU and BertScore for non-OCR parts. For misclassified samples, the cosine similarity, Rouge-L, BELU and BertScore are set directly to 0 as the gist is opposite. Specifically, we extract the content within the quotation marks from the first predicted sentence and use it to calculate the OCR accuracy $Acc_{OCR}$. We then remove stop-words and the content within the quotation marks from both prediction $P_{pred}$ and ground-truth paragraphs $P_{gt}$ to calculate non-OCR scores. The common accuracy metric (Guillaro et al., 2023) is adopted for image forgery classification task.

### 5.2 Implement Details

The vision tower and projector of the large multimodal model are full-parameter fine-tuned and the large language model part is LoRA (Hu et al., 2021) fine-tuned with rank 8 and alpha 16. We adopt AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate decaying linearly from 1e-4 to 0. The batch-size is set to 16 for all models and the experiments are run on NVIDIA A100 80GB GPUs. We set the input area to 1344*896 for the Qwen2-VL and InternVL2 models. In the proposed TextSleuth model, the Swin-Transformer based detection model is trained for 30 epochs on

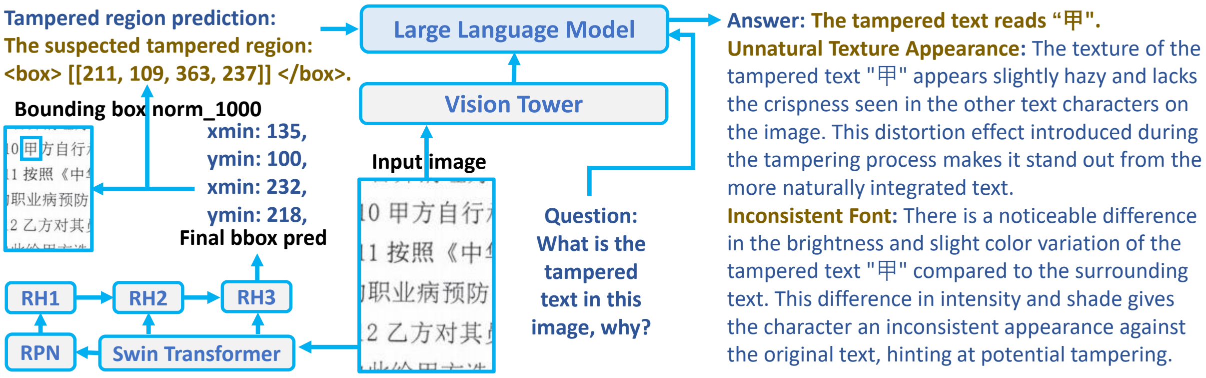



Figure 4: The overall pipeline of the TextSleuth. 'RH' is Refine Head and 'RPN' is Region Proposal Network.

the ETTD-Train, with a batch-size of 16 and an input resolution of 1344*896. The AdamW optimizer is employed with a learning rate that decays linearly from 6e-6 to 3e-6.

For all the fine-tuned models, the input text query is "What is the tampered text in this image, why?", which matches the training data. The image classification prediction is regarded as "tampered" if the edit distance between the model output and the string "There is no tampered text in this image." is greater than 3. For the GPT-4o and other pre-trained models, to output the most similar format with the annotation, the query is set to "Does this image have tampered text on it? Please start your answer with "Yes" or "No". If "Yes", then recognize the tampered text and describe the anomaly of the tampered region". The image classification prediction is regarded as "tampered" if the output starts with "Yes".

### 5.3 Comparison Study

**Anomaly Description**. The comparison results of anomaly description between different LLMs on the ETTD dataset are shown in Table 2. Three conclusions can be drawn through analysis:

(1) **The scaling law applies to the explainable tampered text detection task.** Even within the same series (e.g. InternVL2 or Qwen2-VL) where the vision tower is the same and the pre-training data is similar, models with larger LLMs **mostly** perform better. For example, Qwen2-VL-7B achieves an average final score of 74.2, which is better than Qwen2-VL-2B of 71.6. This confirms that the scaling law behind our task.

(2) **The image artifacts of tampered text are shared across different languages.** When evaluated on ETTD-ML, the accuracy of OCR for tampered text drops significantly for all models except GPT-4o, Qwen2-VL, and TextSleuth. This is because most of the evaluated multimodal LLMs do not naturally learn Arabic, Bengali, Japanese, or Korean well. However, the anomaly description scores have much less degradation compared to the OCR accuracy scores. This is because image artifacts of tampered text are shared across languages, so the models can correctly identify and describe the artifacts in foreign languages. Additionally, our TextSleuth demonstrates consistent performance on both ETTD-ML and ETTD-Test, showing that it can easily transfer its artifact detection ability to most foreign languages.

(3) **The proposed TextSleuth performs the best on both ETTD-Test, ETTD-CD and ETTD-ML, significantly outperforming other models in both in-domain and cross-domain scenarios.** This verifies that the proposed auxiliary prompt improves model's fine-grained perception and helps the model to produce high quality anomaly description by focusing its attention on the suspected tampered region at start.

**Image Forgery Classification**. The comparison results of the image forgery classification are shown in Tab. 4. The public Tampered IC-13 dataset used in evaluation consists of texts tampered by SR-Net (Wu et al., 2019) and can also evaluate model's cross domain performance on unknown tampering method. The proposed TextSleuth considerably outperforms existing methods on all the three benchmarks, and improves the Qwen2-VL-7B baseline by +3.5 points, +7.9 points, +6.6 points and + 12.7 points on the four benchmarks respectively, demonstrating strong effectiveness, the precision and recall scores are shown in Table 7. Similar conclusions can be drawn as analyzed above.

**Auto-annotation**. To verify the effectiveness of the proposed fused mask prompt, we randomly sample 100 tampered text images from the collected data and manually obtain the OCR results of the tampered text. To evaluate the quality of the anomaly description, we further recruit volunteers to score the anomaly descriptions from 0 to 100, where 100 represents perfectly accurate and comprehensive and 0 is the opposite. We compare the average score of both tampered text OCR and anomaly description quality between the binary mask prompt as in existing work (Xu et al., 2024) and the proposed fused mask prompt, the results are shown in Table 5. The annotator GPT-4o has significantly higher OCR accuracy and anomaly description quality with the proposed prompt. This indicates that the GPT-4o with the proposed prompt can better understand the actual location of the tampered text, and therefore can produce more satisfactory anomaly descriptions.

### 5.4 Ablation Study

The ablation study of the proposed TextSleuth is shown in Table 3. We conduct experiments on three base multimodal LLMs, including InternVL2-2B, Qwen2-VL-2B and Qwen2-VL-7B. For each base MLLM, there are four ablation settings. Setting (1) is the official pre-trained model performance. Setting (2) is the official model fine-tuned on the ETTD-Train, which means an end-to-end solution. Setting (3) is the model with our two-stage TextSleuth method and fine-tuned on ETTD-Train. Setting (4) is the model with our TextSleuth method and with the perfect tampered text detector, which is achieved by replacing the predicted tampered text coordinates with the ground-truth coordinates. Three conclusions can be drawn through analysis:
(1) **The existing multimodal models do not have the ability to recognize tampered text and the anomaly.** All three base models perform poorly in setting (1), but much better in setting (2). This confirms that the official open-source models are mostly incapable of detecting the tampered text in our dataset. Training the models on the ETTD data is essential for them to obtain explainable tampered text detection ability.
(2) **The proposed auxiliary prompt can significantly improve model performance across different base models.** For each of the three base models, the model performance in setting (3) is significantly better than that in setting (2) (+11.2 points average final score for InternVL2-2B, +7.6 for Qwen2-VL-2B and +5.4 for Qwen2-VL-7B). These improvements are achieved by the proposed two-stage analysis paradigm and the auxiliary prompt in our TextSleuth. The proposed methods alleviate the difficulty in detecting tampered region and make the models better focused on analyzing the anomaly, resulting in an improved anomaly description quality. Additionally, by focusing on the tampered text with the proposed prompt, the models can learn more general features by reducing the interference from unrelated background styles. Consequently, the model's cross-domain generalization is considerably improved. The huge improvements on different basic multimodal LLMs also demonstrate that our TextSleuth is widely applicable.
(3) **The performance of our TextSleuth can be further improved with better tampered region detectors.** For all of the three base models, model performance in setting (4) is better than those in setting (3). The improvement is achieved by eliminating the errors of the initial tampered text box prediction. Therefore, our TextSleuth can easily be improved in the future with an advanced tampered text region detector.

**Robustness Evaluation**. We evaluate the robustness of the TextSleuth under different JPEG compression quality factors and different resize factors (e.g. resize both image height and width to 0.25 times of the original ones and then resize them to model input resolution) on ETTD-Test, ETTD-CD and ETTD-ML. As shown in Table 6, **the stable average performance has verified the robustness of our TextSleuth on low-quality images**.

More discussions and visualization are presented and analyzed in the Appendix.

## 6 Conclusion

This work pioneers explainable tampered text detection through describing the anomalies of tampered text images in natural language. We propose ETTD, a large-scale comprehensive dataset that consists of multilingual document and scene text images tampered by various methods. We generate anomaly descriptions for the tampered images by prompting GPT-4o with an elaborate query, which effectively instruct GPT-4o to generate comprehensive analysis. Moreover, a fused mask prompt is proposed to more clearly indicate the tampered region for GPT-4o, which significantly reduces confusion and improves the annotation quality. We also propose to filter out the responses with low tampered text OCR accuracy, which can further improve annotation quality in an automatic manner. In addition, we propose a novel model TextSleuth to improve explainable tampered text detection,

| Methods | ETTD-Test (in-domain) | | | | | | ETTD-CD (out-of-domain) | | | | | | ETTD-ML | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OCR | C.S. | BLEU | R.L. | B.S. | Avg. | OCR | C.S. | BLEU | R.L. | B.S. | Avg. | OCR | A.D | Avg. | Avg. |
| GPT-4o (zero-shot) | 48.3 | 66.1 | 12.4 | 21.3 | 59.3 | 41.5 | 74.6 | 78.0 | 17.1 | 27.3 | 70.6 | 53.5 | 62.8 | 48.0 | 51.0 | 48.7 |
| Yi-VL-6B | 49.9 | 76.5 | 44.0 | 50.2 | 73.0 | 58.7 | 64.3 | 81.4 | 34.7 | 43.7 | 76.4 | 60.1 | 32.3 | 58.1 | 52.9 | 57.3 |
| DeepSeekVL-7B | 66.6 | 86.9 | 49.3 | 56.7 | 83.1 | 68.5 | 37.9 | 64.7 | 2.5 | 10.4 | 55.8 | 34.3 | 43.1 | 69.1 | 63.9 | 55.6 |
| MiniCPMV-2.5-8B | 79.3 | 92.6 | 49.0 | 60.5 | 88.1 | 73.9 | 68.9 | 74.8 | 24.3 | 34.5 | 69.8 | 54.5 | 49.0 | 66.9 | 63.3 | 63.9 |
| InternVL2-1B | 77.8 | 89.1 | 48.1 | 56.4 | 85.0 | 71.3 | 79.2 | 84.0 | 33.8 | 43.8 | 79.0 | 64.0 | 48.7 | 59.5 | 57.3 | 64.2 |
| InternVL2-2B | 81.1 | 91.5 | 49.7 | 58.2 | 87.3 | 73.6 | 78.2 | 82.7 | 32.5 | 42.8 | 77.7 | 62.8 | 48.8 | 61.0 | 58.6 | 65.0 |
| InternVL2-4B | 75.8 | 82.4 | 46.4 | 55.0 | 81.6 | 68.2 | 91.4 | 94.0 | 42.3 | 51.5 | 87.8 | 73.4 | 48.5 | 61.9 | 59.2 | 67.0 |
| InternVL2-8B | 80.9 | 90.7 | 49.4 | 57.5 | 86.4 | 73.0 | 80.0 | 85.1 | 35.6 | 45.2 | 80.1 | 65.2 | 50.3 | 60.4 | 58.4 | 65.5 |
| InternVL2-26B | 85.9 | 93.4 | 50.6 | 59.6 | 89.8 | 75.9 | 86.8 | 87.1 | 36.3 | 46.4 | 82.0 | 67.7 | 52.2 | 62.8 | 60.7 | 68.1 |
| Qwen2-VL-2B | 84.8 | 93.7 | 50.5 | 59.4 | 89.4 | 75.6 | 82.1 | 85.0 | 34.8 | 44.7 | 80.1 | 65.3 | 75.0 | 73.6 | 73.9 | 71.6 |
| Qwen2-VL-7B | 87.1 | 94.8 | 50.8 | 59.8 | 90.4 | 76.6 | 87.1 | 89.9 | 39.5 | 49.6 | 84.9 | 70.2 | 85.6 | 73.2 | 75.7 | 74.2 |
| TextSleuth-7B (Ours) | **92.6** | **98.3** | **51.7** | **60.5** | **93.6** | **79.3** | **97.7** | **98.1** | **48.1** | **57.8** | **93.2** | **79.0** | **91.0** | **77.7** | **80.4** | **79.6** |

Table 2: Comparison study of the proposed method. ‘OCR‘ denotes OCR accuracy, ‘C.S.‘ denotes cosine similarity, ‘R.L‘ denotes Rouge-L, ‘B.S‘ denotes Bert score, ‘Avg.‘ denotes average score. ‘A.D‘ denotes the average score of anomaly description, it is calculated by averaging the cosine similarity, BELU, Rouge-L and Bert scores.

| Base MLLM | Ablation Settings | | | | ETTD-Test | | | ETTD-CD | | | ETTD-ML | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Num | SFT | +Ours | +Ours* | OCR | A.D. | Avg. | OCR | A.D. | Avg. | OCR | A.D. | Avg. | Avg. |
| InternVL2-2B | (1) | × | × | × | 14.1 | 29.9 | 26.7 | 34.3 | 33.4 | 33.6 | 5.8 | 26.9 | 22.7 | 27.7 |
| | (2) | ✓ | × | × | 81.1 | 71.7 | 73.6 | 78.2 | 58.9 | 62.8 | 48.8 | 61.0 | 58.6 | 65.0 |
| | (3) | ✓ | ✓ | × | 83.1 | 75.9 | 77.3 | 96.5 | 74.2 | 78.7 | 51.9 | 77.6 | 72.5 | 76.2 |
| | (4) | ✓ | ✓ | ✓ | 83.7 | 76.7 | 78.1 | 97.7 | 75.4 | 79.9 | 52.4 | 78.4 | 73.2 | 77.0 |
| Qwen2-VL-2B | (1) | × | × | × | 18.5 | 29.3 | 27.1 | 29.8 | 32.9 | 32.3 | 19.7 | 29.7 | 27.7 | 29.0 |
| | (2) | ✓ | × | × | 84.8 | 73.3 | 75.6 | 82.1 | 61.2 | 65.4 | 75.0 | 73.6 | 73.9 | 71.6 |
| | (3) | ✓ | ✓ | × | 90.4 | 75.9 | 78.8 | 97.2 | 74.3 | 78.9 | 89.8 | 77.6 | 80.0 | 79.2 |
| | (4) | ✓ | ✓ | ✓ | 91.3 | 76.7 | 79.6 | 98.5 | 75.4 | 80.0 | 91.3 | 78.5 | 81.1 | 80.2 |
| Qwen2-VL-7B | (1) | × | × | × | 14.0 | 23.6 | 21.7 | 36.4 | 28.8 | 30.3 | 16.9 | 29.5 | 27.0 | 26.3 |
| | (2) | ✓ | × | × | 87.1 | 74.0 | 76.6 | 87.1 | 66.0 | 70.2 | 85.6 | 73.2 | 75.7 | 74.2 |
| | (3) | ✓ | ✓ | × | 92.6 | 76.0 | 79.3 | 97.7 | 74.3 | 79.0 | 91.1 | 77.7 | 80.4 | 79.6 |
| | (4) | ✓ | ✓ | ✓ | 93.6 | 76.8 | 80.2 | 99.0 | 75.5 | 80.2 | 92.5 | 78.6 | 81.4 | 80.6 |

Table 3: Ablation study of the proposed method. "SFT" denotes surprised fine-tuning. "+Ours" denotes equipping the model with our proposed TextSleuth method. "+Ours*" denotes using ground-truth tampered region boxes in the TextSleuth's auxiliary prompt. ‘A.D‘ denotes the average score of anomaly description, it is calculated by averaging the cosine similarity, BELU, Rouge-L and Bert scores. ‘Avg.‘ denotes weighted average score.

| Methods | ETTD-Test | ETTD-CD | ETTD-ML | Tamp-IC13 | Average |
|---|---|---|---|---|---|
| GPT-4o (zero-shot) | 67.3 | 79.3 | 73.7 | 82.8 | 75.8 |
| Yi-VL-6B | 76.9 | 81.9 | 74.0 | 45.9 | 69.7 |
| DeepSeekVL-7B | 87.4 | 66.7 | 86.5 | 76.4 | 79.3 |
| MiniCPMV-2.5-8B | 93.2 | 75.5 | 83.7 | 56.7 | 77.3 |
| InternVL2-1B | 89.7 | 84.6 | 71.3 | 59.2 | 76.2 |
| InternVL2-2B | 92.1 | 83.3 | 73.6 | 58.8 | 77.0 |
| InternVL2-4B | 82.8 | 94.5 | 75.4 | 36.1 | 72.2 |
| InternVL2-8B | 91.2 | 85.7 | 72.4 | 60.5 | 77.5 |
| InternVL2-26B | 92.4 | 86.1 | 76.7 | 63.2 | 79.6 |
| Qwen2-VL-2B | 94.3 | 85.7 | 93.3 | 73.8 | 86.8 |
| Qwen2-VL-7B | 95.4 | 90.5 | 92.3 | 75.1 | 88.3 |
| TextSleuth-7B (Ours) | **98.9** | **98.6** | **98.9** | **88.4** | **96.2** |

Table 4: Accuracy performance of different large multi-modal models on image forgery classification task.

| Method | OCR Accuracy | Perfect Match | Quality Score |
|---|---|---|---|
| Binary mask prompt | 47.3 | 30.4 | 63.2 |
| Fused mask prompt (Ours) | **84.2** | **73.0** | **85.7** |

Table 5: Comparison for the fused mask prompt.

| All avg. score | Ori. | JPEG compress quality75 | JPEG compress quality50 | Image resize factor0.75 | Image resize factor0.25 |
|---|---|---|---|---|---|
| Qwen2-VL | 74.2 | 73.0 | 71.8 | 73.6 | 72.7 |
| TextSleuth | **79.6** | **78.8** | **78.1** | **79.3** | **78.7** |

Table 6: Robustness evaluation.

| Models | ETTD-Test | | ETTD-CD | | ETTD-ML | | T-IC13 | |
|---|---|---|---|---|---|---|---|---|
| | P | R | P | R | P | R | P | R |
| Qwen2-VL-7B | .996 | .927 | .881 | .992 | .989 | .878 | .885 | .775 |
| Ours-7B | **.997** | **.985** | **.985** | **.994** | **.992** | **.986** | **.975** | **.893** |

Table 7: Precision (P) and recall (R) scores on image forgery classification task.

which overcomes several major challenges in the field with a two-stage analysis paradigm and an auxiliary prompt. Experiments have confirmed that the proposed method considerably improves upon different baseline models, and that our TextSleuth notably outperforms existing multimodal large language models in both in-domain and cross-domain evaluation on both the ETTD and public datasets. We believe that our valuable ETTD dataset and our first-of-its-kind, simple-yet-effective methods can shed light on the further research in this field.

**Limitations**. Despite the fact that our TextSleuth brings significant improvements to different base multimodal LLMs. Our two-stage paradigm may introduce error accumulation in a very small number of samples. Additionally, the first stage's tampered region prediction may be inconsistent with the second stage's textual prediction.

## References


Amr Gamal Hamed Ahmed and Faisal Shafait. 2014. Forgery detection based on intrinsic document contents. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 252–256. IEEE.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang Wang, et al. 2024a. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Renshuai Liu Dong, Li, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. 2024. Robust text image tampering localization via forgery traces enhancement and multiscale attention. *IEEE Transactions on Consumer Electronics*.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. 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*, pages 20606–20615.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, Wenming Yang, and Jiaya Jia. 2024. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *Preprint*, arXiv:2408.10072.

Christoph H Lampert, Lin Mei, and Thomas M Breuel. 2006. Printing technique classification for document counterfeit detection. In *2006 International Conference on Computational Intelligence and Security*, volume 1, pages 639–644. IEEE.

Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. 2024. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *Preprint*, arXiv:2410.10238.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. 2024. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, page 110828.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2023. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582.

Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. 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*, pages 5937–5946.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. 2024a. Generalized tampered scene text detection in the era of generative ai. *Preprint*, arXiv:2407.21422.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. 2024b. Omni-iml: Towards unified image manipulation localization. *Preprint*, arXiv:2411.14823.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Huiru Shao, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. 2023. Progressive supervision for tampering localization in document images. In *International Conference on Neural Information Processing*, pages 140–151. Springer.

Huiru Shao, Zhuang Qian, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. 2025. Delving into adversarial robustness on document tampering localization. In *European Conference on Computer Vision*, pages 290–306. Springer.

Yalin Song, Wenbin Jiang, Xiuli Chai, Zhihua Gan, Mengyuan Zhou, and Lei Chen. 2024. Cross-attention based two-branch networks for document image forgery localization in the metaverse. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024. Forgerysleuth: Empowering multimodal large language models for image manipulation detection. *Preprint*, arXiv:2411.19466.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

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

Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. 2019. Editing text in the wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1500–1508, New York, NY, USA. Association for Computing Machinery.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2024. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *Preprint*, arXiv:2410.02761.

Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. 2019. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE.

# TextSleuth: Towards Explainable Tampered Text Detection

Anonymous ACL submission


## Abstract

In this supplementary material, we show our detailed textual prompt that is elaborately designed to guide the GPT-4o to describe the anomaly of the manipulated text. Moreover, we present more discussions about this work. In addition, we show the prediction of GPT-4o, Qwen2VL-7B and our TextSleuth for visual comparison. Finally, we present more examples and their annotations in the proposed ETTD dataset.


## 1 The Proposed Textual Prompt

Due to the different characteristics of tampered text, existing textual prompts designed for natural objects or deepfakes cannot be directly used to generate high-quality anomaly descriptions for tampered text. To this end, we redesign the textual prompt by summarizing the possible anomalies caused by text tampering into six major perspectives and providing a detailed explanation for each of them.

The full version of our textual prompt is:

You are an expert good at analyzing tampered text images. You will be provided with two images, **the first is the tampered text image A and the second is the reference image B, with the tampered areas highlighted and the authentic areas darkened**.

Your task is to: **First, recognize the tampered text and output its OCR result. Second, Describe visible details in the image that have been tampered with.** Please consider the visible details caused by tampering from these perspectives.

**1. Edge artifacts**. The background of the tampered text may be inconsistent with the authentic regions. Therefore, the edges around the tampered text region may be discontinuous and inconsistent with the background.

**2. Unnatural texture appearance**. The texture appearance of the tampered text may be slightly blurred, hazy, jagged, have a distortion effect, or have an unnatural clarity.

**3. Inconsistent font**. The font of the tampered text may be slightly different in color, size, thickness, brightness, or style from the surrounding real text.

**4. Inconsistent alignment**. The tampered text may have inconsistent spacing with the surrounding text or a small offset to the text line.

**5. Text incoherence**. Tampered text may break the coherence of the sentence.

**6. Lack of integration**. The tampered text may appear unnaturally placed and not integrated with its surroundings, or it may not blend seamlessly with its surroundings, appearing artificially overlaid or unnaturally pasted. Don't mention the image B in your answer, always assume that you are only observing the input image A.

As shown in Figure 1, our proposed prompt can help GPT-4o output comprehensive and accurate anomaly descriptions.

## 2 Discussions

**1. Why use a two-stage paradigm instead of end-to-end MLLM for tampered text detection.** MLLM are well known to perform poorly in predicting coordinates. For example, Swin-large achieves 57.7 AP on COCO but Qwen2-VL 7B merely achieves 43.9. We fine-tune Intern2VL-2B and Qwen2VL-7B with their official pipelines on our dataset, their poor detection performance in the table below again confirms this conclusion. This is the reason why we adopts a two-stage paradigm instead of an end-to-end MLLM for detection.

**2. Is it possible for GPT-4o to generate misleading descriptions? If so, how to avoid this?** Our annotation process was rigorously designed and iteratively optimized to minimize discrepancies. The GPT-4o anomaly description for each instance is mostly within 6 perspectives (Figure 1 of the Appendix). Since the annotations of the test data are manually washed, we evaluate the discrepancy by calculating the proportion of description elements that are deleted, added or significantly revised. The result is 5.87%, which means that the discrepancy between GPT-4o and human annotations is small. Although our method effectively reduces error descriptions, some occasional errors are unavoidable. To tackle this, we manually check and correct the wrong descriptions for the test data. But washing the training data is costly, and a tiny proportion of noisy data (merely 5.87% imperfect) is common in real-world scenarios. So we just keep them, and thus our dataset can better reflect the model robustness in the real world.

**3. Why not use traditional multimodal methods such as BLIP?** BLIP and LLaVA are only for English and have too small fixed input shapes (224x224 and 336x336), resizing images to such a small size will make most of the artifacts invisible, so they are not suitable for this task.

**4. Semantic analysis and reasoning.** Both our dataset and our method have already incorporated semantic analysis as a core component. For example, in the upper middle of Figure 1 in the Appendix, we prompt the annotator model to pay special attention to semantic analysis (titled "5. Text incoherence" in our proposed prompt). Accordingly, in the lower middle of Figure 1 in the appendix, the model successfully analyzes the semantic error (should be "BREAK GLASS" instead of "BREAK GRASS" in safety instructions). With such sufficient training data for semantic analysis, our trained model naturally incorporates semantic analysis and reasoning as a core component in its prediction.

**5. The source images in our ETTD dataset are not based on outdated datasets.** First, we do not rely entirely on the existing datasets. To further increase the diversity and challenge of our dataset, we have utilized thousands of recent challenging text images from the Internet, as shown in the 3rd column of Table 1 in paper. This ensures that our dataset is both up-to-date and challenging. Second, the ICDAR2013, ICDAR2017-MLT and LSVT2019 datasets remain highly relevant and valuable resources. They are widely recognized in current research as qualified datasets for creating high-quality tampered text (Wang et al., 2022; Qu et al., 2025; Shu et al., 2024, 2025), making them appropriate choices for our work.

**6. ICDAR has multiple datasets and competition benchmarks, specify which task and dataset are being referred to.** Although different tasks in an ICDAR dataset have different annotations, they share the same dataset images. We specify the task and url as following: ICDAR2013: Task 2.1: Text Localization (https://rrc.cvc.uab.es/?ch=2&com=downloads). ICDAR2017: Task 1: Multi-script text detection (https://rrc.cvc.uab.es/?ch=8&com=downloads). LSVT: The training set with fully annotated images. (https://rrc.cvc.uab.es/?ch=16&com=downloads)

**7. Lower image quality affects model performance?** Lower image quality will undoubtedly cause performance degradation to all of the forensic models (no exception) (Guillaro et al., 2023; Liu et al., 2022; Qu et al., 2025). However, as analyzed in Lines 562 to 570 of the paper, our model has minimized performance degradation under various distortions and shows strong robustness to lower image quality. For small size images, we can simply resize them to our model input resolution (1344*896) and they are just more blurry compared to high-quality images. Our model is almost unaffected by this process (Table 6 of paper).

**8. How does this work benefit the ACL community?** This research directly aligns with ACL's interest in multimodal language understanding and vision-language integration, which are growing areas of focus within computational linguistics. This work focuses on improving model ability to linguistically explain text forgery, which is **fundamentally a language-based task** requiring sophisticated natural language generation and reasoning. This work is the first work on explainable tampered text detection, proposes novel, practical methods and valuable dataset, successfully extending the capability boundaries of language models.

Table 1: The detection performance of the detector in the proposed TextSleuth. The IoU threshold is 0.5.

| Dataset | Intern2VL-2B | Qwen2VL-7B | Ours |
|---|---|---|---|
| ETTD-Test | 0.705 | 0.743 | **0.990** |
| ETTD-CD | 0.665 | 0.692 | **0.989** |

## 3 Visualization

The prediction visualization of GPT-4o, Qwen2-VL-7B and our TextSleuth is shown in Figures 2 and 3. Evidently, the proposed method can produce more accurate results. More samples of the ETTD dataset are shown in Figures 4 and 5, demonstrating the variety of the proposed dataset.

| Language | OCR | Cosine | Rouge-L | BELU | BertScore |
|---|---|---|---|---|---|
| Arabic | 90.8 | 98.1 | 67.0 | 59.2 | 93.8 |
| Bangla | 84.7 | 98.5 | 63.0 | 54.3 | 93.5 |
| Japanese | 96.0 | 98.2 | 62.9 | 53.9 | 93.7 |
| Korean | 92.6 | 97.7 | 62.7 | 53.4 | 93.1 |

Table 2: The detailed performance of our TextSleuth on ETTD-ML.

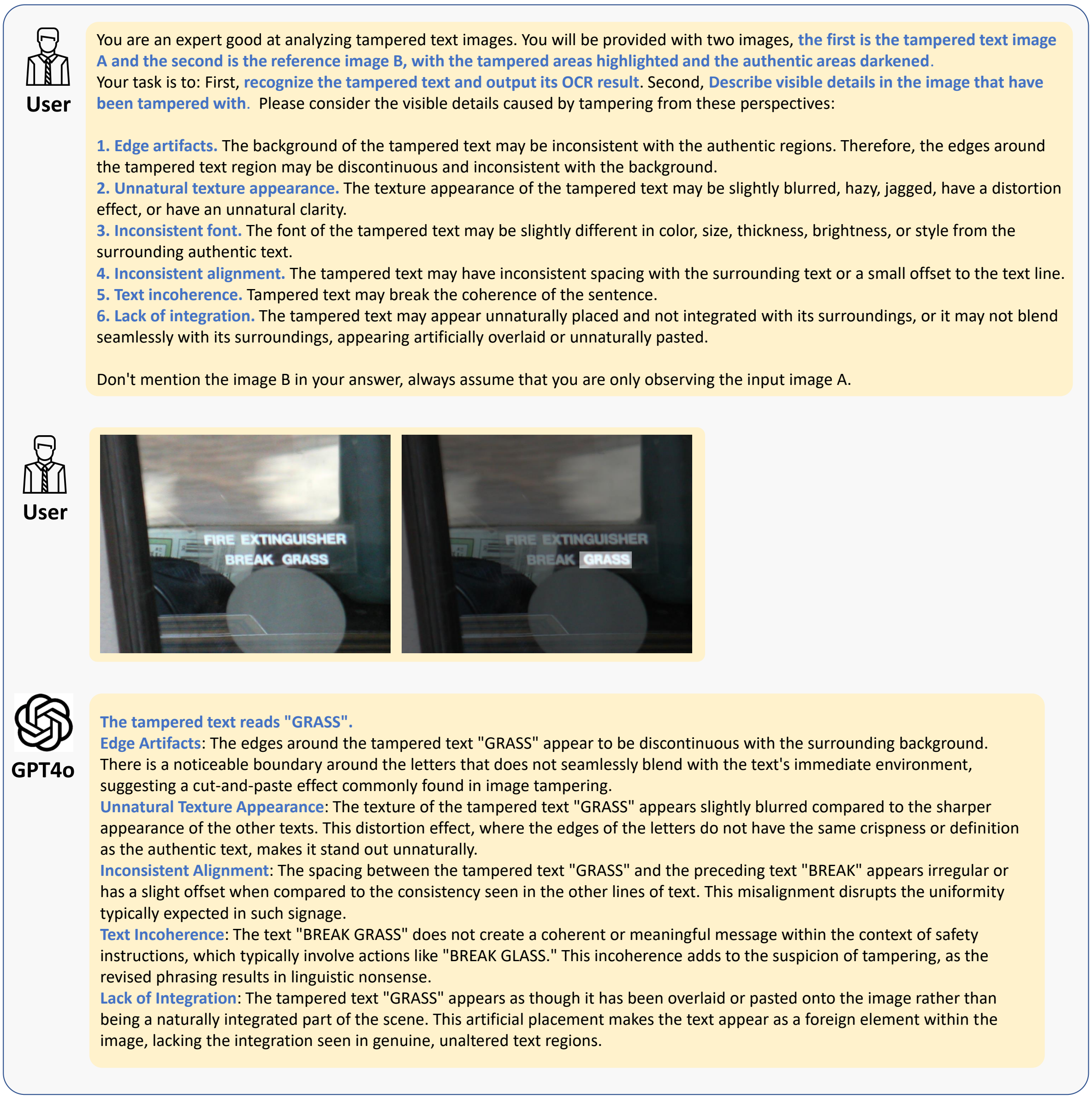



Figure 1: Our proposed textual prompt are specially designed for tampered text can can guide GPT-4o to generate high-quality anomaly descriptions for tampered text.

## References

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. 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*, pages 20606–20615.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 2022. 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.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. 2025. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol-

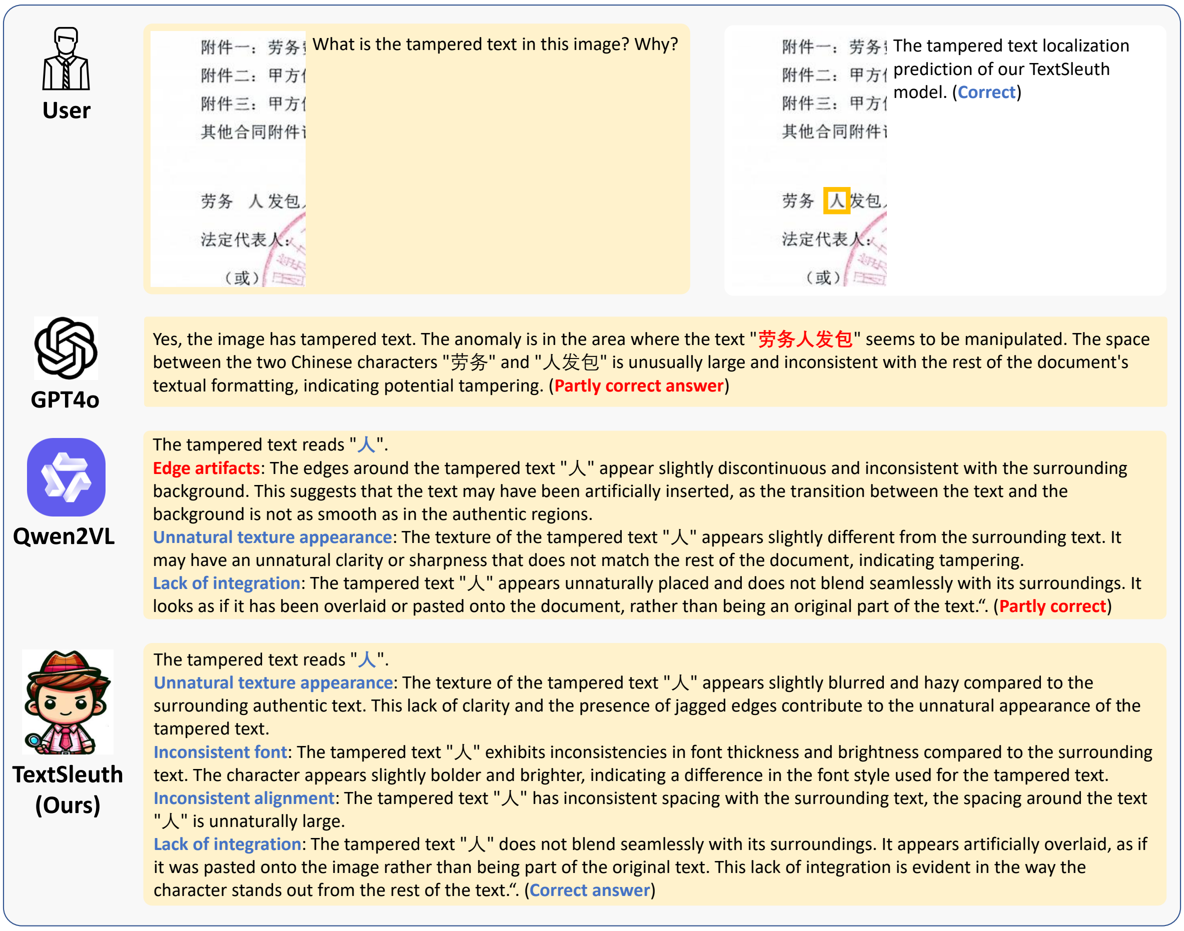



Figure 2: Our proposed TextSleuth can accurately detect the tampered text and better explain the anomalies of the tampered text.

ume 39, pages 694–702.

Yan Shu, Weichao Zeng, Zhenhang Li, Fangmin Zhao, and Yu Zhou. 2024. Visual text meets low-level vision: A comprehensive survey on visual text processing. *arXiv preprint arXiv:2402.03082*.

Yan Shu, Weichao Zeng, Fangmin Zhao, Zeyu Chen, Zhenhang Li, Xiaomeng Yang, Yu Zhou, Paolo Rota, Xiang Bai, Lianwen Jin, et al. 2025. Visual text processing: A comprehensive review and unified evaluation. *arXiv preprint arXiv:2504.21682*.

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

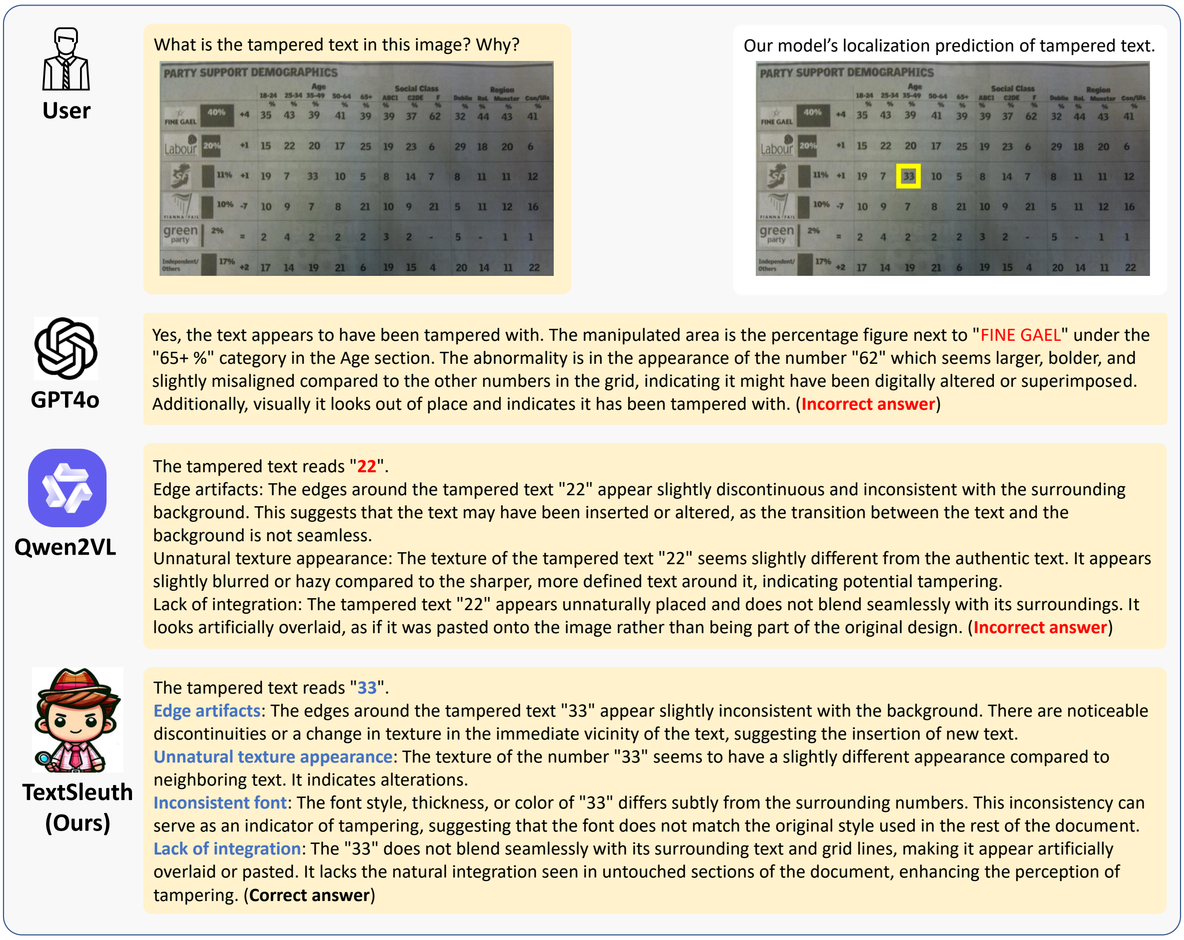
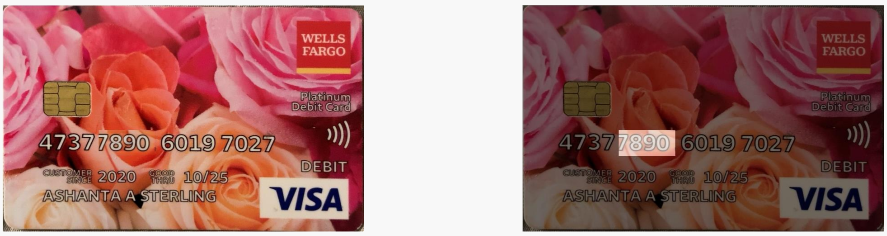



Figure 3: Our proposed TextSleuth can accurately detect the tampered text and better explain the anomalies of the tampered text.


**The tampered text reads "PR".**
**Edge Artifacts**: The tampered text "PR" shows visible discontinuities at the edges compared to the surrounding areas. The edges around the "PR" text show a break or discontinuity against the background, indicating it was not part of the original image.
**Unnatural Texture Appearance**: The texture of the tampered "PR" text appears somewhat hazy and less sharp than the surrounding authentic texts. This indicates manipulation, as the integrated texture should normally appear consistent.
**Inconsistent Font**: There is a noticeable inconsistency in size and color with the text just before it and the authentic "STOP" text below. The tampered text "PR" appears slightly more faded and not as bold as the surrounding text.
**Inconsistent Alignment**: The "PR" does not line up properly with the preceding text "DON," appearing offset both horizontally and vertically. This inconsistent alignment makes the text look less natural and more likely tampered with.
**Text Incoherence**: The inclusion of the letters "PR" within the context of the image does not form a coherent or expected phrase within the setting of a traffic sign, suggesting tampering in terms of intended message.
**Lack of Integration**: The tampered "PR" text does not smoothly integrate with the color and shading nuances of the stop sign's red background or the adjacent legitimate text, making it appear as an artificially placed element.


**The tampered text reads "JAMES".**
**Edge artifacts**: The edges around the tampered text "JAMES" appear to be inconsistent with the background, indicating discontinuity. This is evident where the boundary between the tampered region and the surrounding area is not as smooth or consistent as the original regions.
**Unnatural texture appearance**: The tampered text "JAMES" exhibits a slightly unnatural texture, where the clarity of the letters seems detached from the rest of the image. This could manifest as either an overly sharp appearance compared to the slightly blurred background or an inconsistent texture.
**Inconsistent font**: There is a noticeable difference in the font style or thickness of the tampered text "JAMES" compared to the surrounding authentic text. This can include variations in color, size, or brightness, which makes the tampered text stand out unnaturally from the rest.
**Lack of integration**: The tampered text "JAMES" appears unnaturally placed, as if overlaid on the image without seamlessly blending with its surroundings. This lack of integration points to an artificial addition which has a contrasting appearance compared to the authentic text.




**The tampered text reads "7890".**
**Edge Artifacts**: The edges around the tampered text "7890" appear to be slightly discontinuous and inconsistent with the background texture, which results in an unnatural demarcation between the tampered text and the surrounding authentic text. This discontinuity suggests tampering.
**Unnatural Texture Appearance**: The tampered text "7890" has a slightly different texture compared to the surrounding numbers. It appears somewhat blurred or artificially smoothened, which results in an unnatural clarity that is distinct from the rest of the text, indicating a potential tampering effect.
**Lack of Integration**: The tampered text "7890" does not blend seamlessly with the surrounding characters and the background. It appears somewhat as if it is placed on top of the card's design rather than integrated into it, giving an impression of being unnaturally overlaid or pasted into the image.

Figure 4: More data samples in the proposed ETTD dataset. The left image in each row is the original image and the tampered text region is highlighted in the right image.



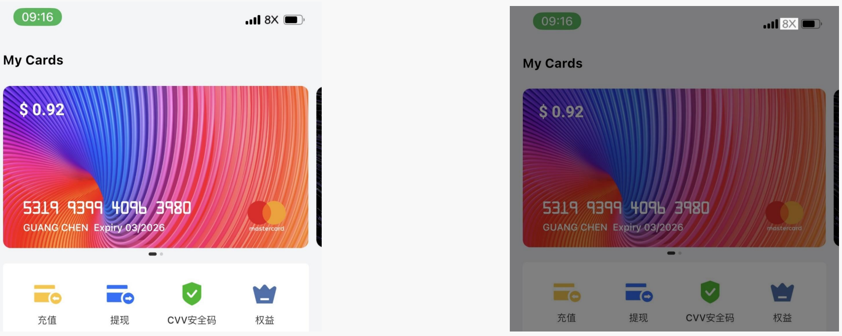



Figure 5: More data samples in the proposed ETTD dataset. The left image in each row is the original image and the tampered text region is highlighted in the right image.