TextSleuth: Towards Explainable Tampered Text Detection

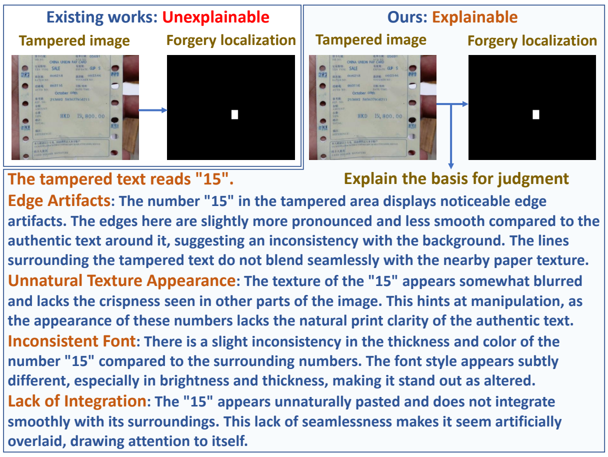
Anonymous ACL submission

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 methods are employed to improve the quality of our dataset, such as using elaborated queries to generate high-quality anomaly descriptions with GPT-4o. A fused mask prompt is proposed to reduce confusion when querying GPT-4o to generate anomaly descriptions. To automatically filter out low-quality annotations, we also propose to prompt GPT-4o to recognize tampered texts before describing the anomaly, and to filter out the responses with low OCR accuracy. To further improve explainable tampered text detection, we propose a simple yet effective model called **TextSleuth**, which improves fine-grained perception and cross-domain generalization by focusing on the suspected region, with a two-stage analysis paradigm and an auxiliary grounding prompt. 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

Text image is one of the most important information carriers in today's society, containing a large amount of sensitive and private information (Chen et al., 2024a). 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). It is crucial to develop effective and reliable methods for detecting tampered text images.



The tampered text reads "15".

Edge Artifacts: The number "15" in the tampered area displays noticeable edge artifacts. The edges here are slightly more pronounced and less smooth compared to the authentic text around it, suggesting an inconsistency with the background. The lines surrounding the tampered text do not blend seamlessly with the nearby paper texture.

Unnatural Texture Appearance: The texture of the "15" appears somewhat blurred and lacks the crispness seen in other parts of the image. This hints at manipulation, as the appearance of these numbers lacks the natural print clarity of the authentic text.

Inconsistent Font: There is a slight inconsistency in the thickness and color of the number "15" compared to the surrounding numbers. The font style appears subtly different, especially in brightness and thickness, making it stand out as altered.

Lack of Integration: The "15" appears unnaturally pasted and does not integrate smoothly with its surroundings. This lack of seamlessness makes it seem artificially overlaid, drawing attention to itself.

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.

Existing works model tampered text detection as semantic segmentation (Shao et al., 2023) or object detection (Qu et al., 2024a), with the aim of interpreting the basis for image forgery classification by predicting tampered regions. Despite the progress made in recent years, such fine-grained predictions are still black-box and cannot provide a convincing basis to support the judgement, leading 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 such as ICDAR2017 (Nayef et al., 2017) and LSVT (Sun et al., 2019). 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., 2024b). Poisson Blending (Pérez et al., 2023) is employed to reduce the visual inconsistency around tampered region. Finally, we create 12,000 tampered text images with accurate pixel-level annotations of the tampered region and 10,500 authentic text images. The large-scale of our data notably alleviates the data hunger of deep models. The images are split to ETTD-Train, ETTD-Test and ETTD-CD, the two test sets have the same and different distributions with the ETTD-Train respectively, allowing both in-domain and cross-domain evaluation.

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 specifying 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 out 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., 2024c), 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. For example, a fused mask prompt to reduce model confusion and a novel method to automatically filter out bad responses. 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 conclusions and insights are provided through in-depth analysis, inspiring the further research in this field.

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 Explainable Tampered Text Detection (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 (e.g. ICDAR2013 (Karatzas et al., 2013) and LSVT (Sun et al., 2019). 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., 2024b), 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,

Table 1: A brief summary of the ETTD dataset statistics. "Forged Area" denotes the area ratio of tampered text.

Dataset	Image types	Image source	Languages	Tampering Types (# of samples)	Authentic 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

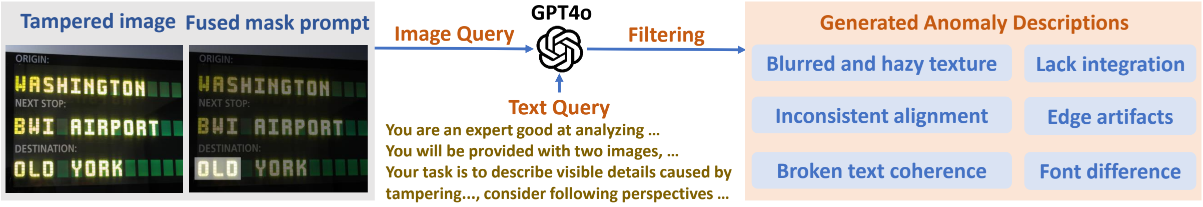


Figure 2: The pipeline for obtaining the textual anomaly description for the tampered text.

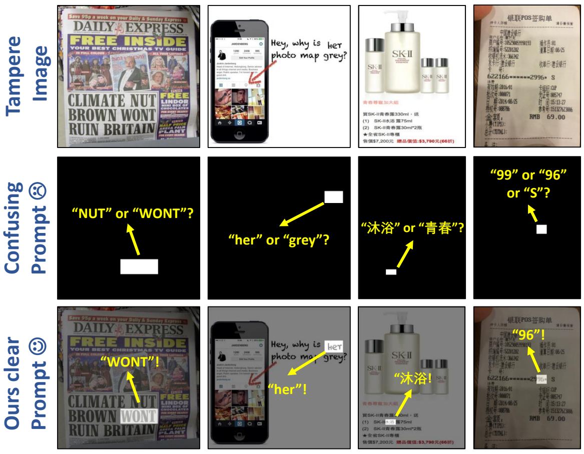


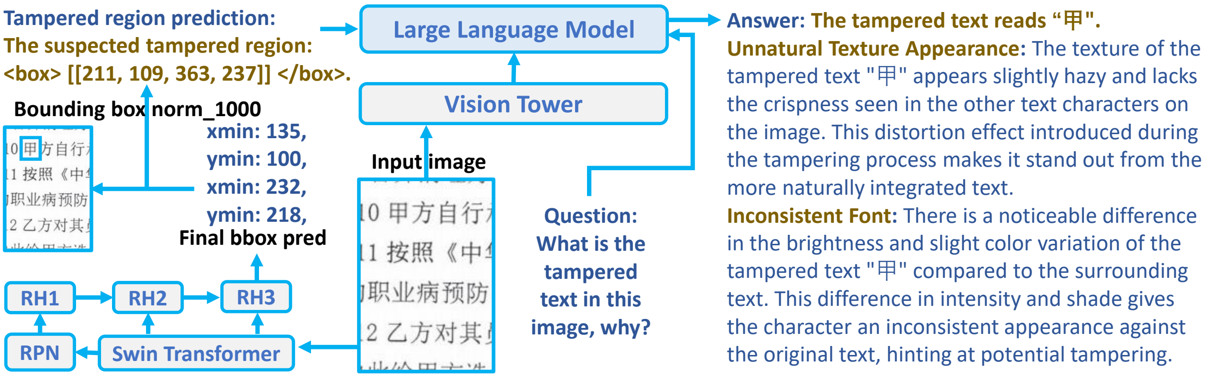
Figure 3: The binary mask prompt as in existing work is confusing in text images. In contrast, our proposed fused mask prompt clearly indicates the content and the exact location of the tampered text.

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 λ_1 and λ_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 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 5,500 text images tampered by copy-move, 5,500 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 10,500 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. Accurate pixel-level annotations for tampered regions are provided to facilitate fine-grained analysis of the tampered text regions.



Tampered region prediction:
The suspected tampered region:
<box> [[211, 109, 363, 237]] </box>.

Question: What is the tampered text in this image, why?

Answer: The tampered text reads "甲".
Unnatural Texture Appearance: The texture of the tampered text "甲" appears slightly hazy and lacks the crispness seen in the other text characters on the image. This distortion effect introduced during the tampering process makes it stand out from the more naturally integrated text.
Inconsistent Font: There is a noticeable difference in the brightness and slight color variation of the tampered text "甲" compared to the surrounding text. This difference in intensity and shade gives the character an inconsistent appearance against the original text, hinting at potential tampering.

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

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., 2024c) 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 paragraph cosine similarity for non-OCR parts. The weighted summary of OCR accuracy and paragraph similarity is used as the final similarity score. For misclassified samples, the paragraph cosine similarity is 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} for more accurate paragraph similarity calculation. The paragraph feature vectors V_{pred}, V_{gt} are obtained by averaging the word vectors in each paragraph, $V_{pred} = average([W2V(word)\ for\ word\ in\ P_{pred}])$, $V_{gt} = average([W2V(word)\ for\ word\ in\ P_{gt}])$, where $W2V$ is the pretrained word-to-vector function (Mikolov et al., 2018). Finally, the cosine similarity between the two paragraph feature vectors is used as the paragraph similarity score, $Sim_{para} = Cos(V_{pred}, V_{gt})$. We have manually verified that better predictions almost always lead to higher cosine similarity scores. The final similarity score Sim is calculated by $Sim = 0.3 * Acc_{OCR} + 0.7 * Sim_{para}$. 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 maximum input area to 1344*896 for the Qwen2-VL models. In the proposed TextSleuth model, the Swin-Transformer based detection model is trained for 30 epochs on the ETTD-Train, with a batch-size of 16 and a maximum 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 pretrained 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) **High input resolution is essential for models to correctly recognize the tampered text and explain its anomaly, but it can also slightly weaken model generalization to unfamiliar scenarios.** On the ETTD-Test set, the Yi-VL-6B with the lowest input resolution 448×448 achieves the lowest final score of 68.5, which is notably lower than other fine-tuned models. Resizing the input image to such a small resolution makes the subtle visual anomaly hard to detect, thus Yi-VL-6B performs the worst. However, the Yi-VL-6B performs much better on ETTD-CD, where the tampered text is mostly larger. Most of the other models suffer significant performance degradation due to the unfamiliar scenario. This indicates that high input resolution mostly weaken model generalization on unfamiliar scenarios.

(2) **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 90.7, which is better than Qwen2-VL-2B. This confirms that the scaling law behind our task.

(3) **The proposed TextSleuth performs the best on both ETTD-Test and ETTD-CD, 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 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 and + 12.7 points on the three benchmarks respectively, demonstrating the effectiveness of the proposed method. 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. For each base model (e.g. InternVL2-2B), 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. Setting (3) is the TextSleuth fine-tuned with the proposed grounding auxiliary prompt. Setting (4) is the TextSleuth 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 tampered text. Training them on the ETTD data is essential for them to achieve explainable tampered text detection.

(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) (+10.8 points average final score for InternVL2-2B, +9.2 for Qwen2-VL-2B and +6.5 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 on ETTD-Test and ETTD-CD. As shown in Table 6, the stable performance under various distortions has verified the robustness of our TextSleuth.

The visualization of the model prediction is 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. To address the lack of dataset, 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. Given that the incorrect recognition of tampered text means unclear perception and leads to bad anomaly description, 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 addi-

Table 2: Comparison study of the proposed method.

Methods	Maximum Input Resolution	ETTD-Test (in-domain) OCR Accuracy	ETTD-Test (in-domain) Cosine Similarity	ETTD-Test (in-domain) Final Score	ETTD-CD (cross-domain) OCR Accuracy	ETTD-CD (cross-domain) Cosine Similarity	ETTD-CD (cross-domain) Final Score	Average Final Score
GPT-4o	-	48.3	66.1	60.7	74.6	78.0	77.0	68.9
Yi-VL-6B	448×448	49.9	76.5	68,5	64.3	81.4	76.2	72.4
DeepSeekVL-7B	1024×1024	66.6	86.9	80.8	37.9	64.7	56.7	68.8
MiniCPMV-2.5-8B	1792×896	79.3	92.6	88.6	68.9	74.8	73.0	80/8
InternVL2-1B	1344×896	77.8	89.1	85.7	79.2	84.0	82.5	84.1
InternVL2-2B	1344×896	81.1	91.5	88.3	78.2	82.7	81.3	84.8
InternVL2-4B	1344×896	75.8	82.4	80.4	91.4	94.0	93.1	86.8
InternVL2-8B	1344×896	80.9	90.7	87.7	80.0	85.1	83.5	85.6
Qwen2-VL-2B	1344×896	84.8	93.7	91.0	82.1	85.0	84.1	87.6
Qwen2-VL-7B	1344×896	87.1	94.8	92.4	87.1	89.9	88.9	90.7
TextSleuth-7B (Ours)	1344×896	**92.6**	**98.3**	**96.5**	**97.7**	**98.1**	**97.9**	**97.2**

Table 3: Ablation study of the proposed method. "SFT" denotes surprised fine-tuning. "TextSleuth" denotes equipping the model with the proposed TextSleuth method. "Perfect Detector" denotes using ground-truth tampered region boxes in the TextSleuth's auxiliary prompt.

Base Multi-modal Model	Ablation settings Num	Ablation settings SFT	Ablation settings Text Sleuth	Ablation settings Perfect Detector	ETTD-Test (in-domain) OCR accuracy	ETTD-Test (in-domain) Cosine similarity	ETTD-Test (in-domain) Final score	ETTD-CD (cross-domain) OCR accuracy	ETTD-CD (cross-domain) Cosine similarity	ETTD-CD (cross-domain) Final score	Average Final score
InternVL2-2B	(1)	×	×	×	14.1	57.8	44.7	34.3	64.4	55.4	50.1
	(2)	✓	×	×	81.1	91.5	88.3	78.2	82.7	81.3	84.8
	(3)	✓	✓	×	83.1	98.3	93.6	96.5	98.1	97.5	95.6
	(4)	✓	✓	✓	83.7	99.4	94.6	97.7	99.4	98.8	96.7
Qwen2-VL-2B	(1)	×	×	×	18.5	57.0	45.5	29.8	63.8	53.6	49.6
	(2)	✓	×	×	84.8	93.7	91.0	82.1	85.0	84.1	87.6
	(3)	✓	✓	×	90.4	98.2	95.8	97.2	98.0	97.7	96.8
	(4)	✓	✓	✓	91.3	99.3	96.8	98.5	99.3	99.0	97.9
Qwen2-VL-7B	(1)	×	×	×	14.0	41.8	33.5	36.4	53.4	48.3	40.9
	(2)	✓	×	×	87.1	94.8	92.4	87.1	89.9	88.9	90.7
	(3)	✓	✓	×	92.6	98.3	96.5	97.7	98.1	97.9	97.2
	(4)	✓	✓	✓	93.6	99.4	97.6	99.0	99.4	99.2	98.4

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

Method	ETTD-Test (in-domain)	ETTD-CD (out-domain)	Tampered-IC13 (zero-shot)
GPT-4o	67.3	79.3	82.8
Yi-VL-6B	76.9	81.9	45.9
DeepSeekVL-7B	87.4	66.7	76.4
MiniCPMV2.5	93.2	75.5	56.7
InternVL2-1B	89.7	84.6	59.2
InternVL2-2B	92.1	83.3	58.8
InternVL2-4B	82.8	94.5	36.1
InternVL2-8B	91.2	85.7	60.5
Qwen2-VL-2B	94.3	85.7	73.8
Qwen2-VL-7B	95.4	90.5	75.1
TextSleuth-7B	**98.9**	**98.6**	**88.4**

Table 5: Comparison for the fused mask prompt.

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 6: Robustness evaluation.

Average final score	Ori.	JPEG compress quality75	JPEG compress quality50	Image resize factor0.75	Image resize factor0.5
Qwen2-VL	90.7	89.6	87.2	89.2	86.1
TextSleuth	**97.2**	**96.3**	**94.4**	**95.8**	**93.0**

tion, we propose a novel model TextSleuth to improve explainable tampered text detection, 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. In-depth analysis is also provided to inspire further work. We believe that our valuable ETTD dataset and our first-of-its-kind, simple-yet-effective methods can shed light on the further research on interpretable tampered text detection.

Limitations. Despite the fact that our TextSleuth performs best and has the minimal performance degradation on the cross-benchmark evaluation, its cross-benchmark performance (e.g. trained on the ETTD-Train and tested on the Tampered IC-13) still has has a lot of room for improvement. We will try to explore how to make the TextSleuth learn more generalized features and further improve its performance on unseen scenarios.

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)*.

Changsheng Chen, Liangwei Lin, Yongqi Chen, Bin Li, Jishen Zeng, and Jiwu Huang. 2024a. Cma: A chromaticity map adapter for robust detection of screen-recapture document images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15577–15586.

Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang Wang, et al. 2024b. 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. 2024c. 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.

Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE.

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)*.

Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. 2017. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE.

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*.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. 2019. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE.

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 also show the performance of the detector in TextSleuth. 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 Detection Performance

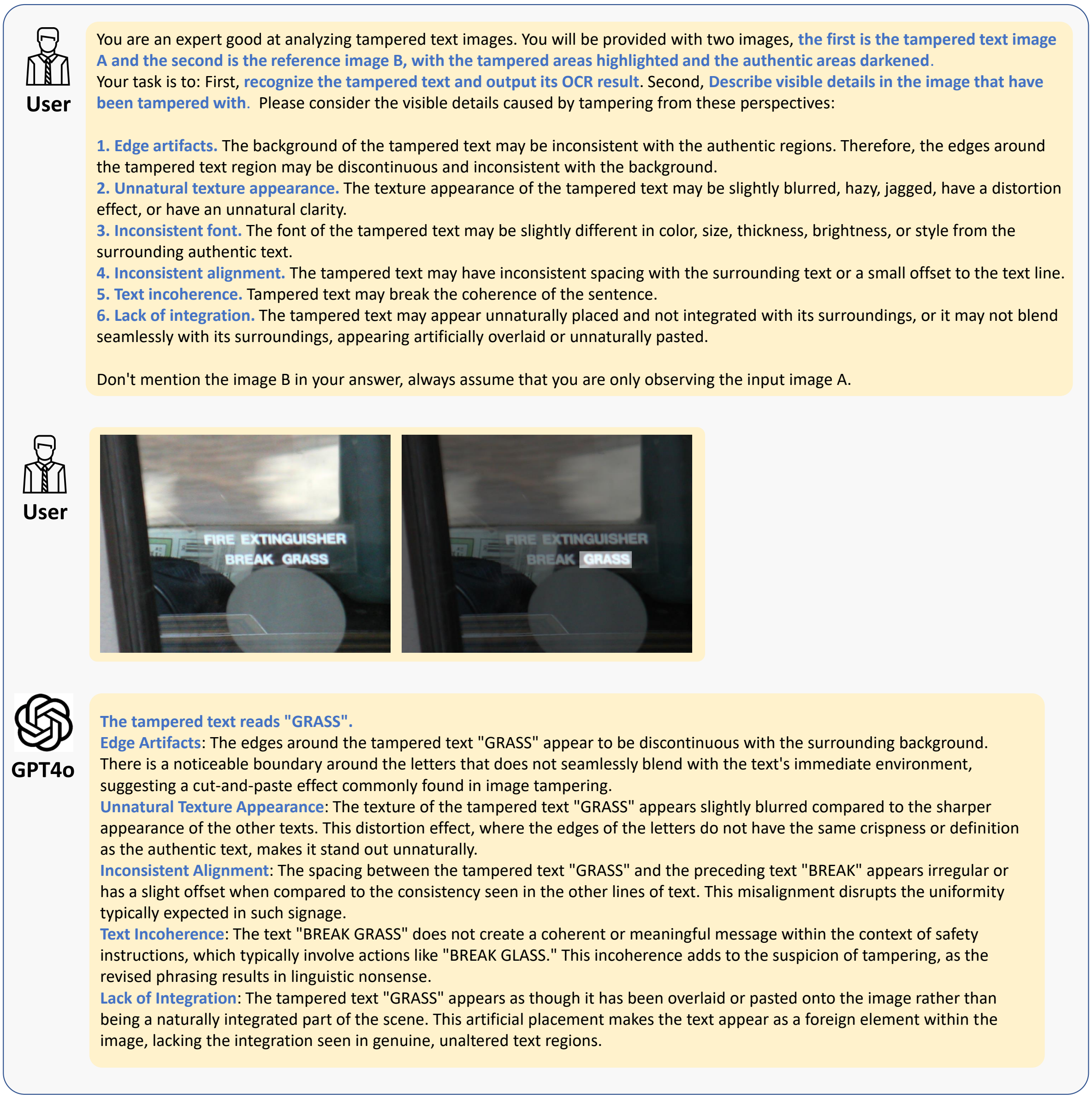
We present the detection performance of the detector in the proposed TextSleuth in Table 1. The precision, recall and F1-score under the ICDAR2017 DetEval protocol and an IoU threshold of 0.5.

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

Dataset	Precision	Recall	F1-score
ETTD-Test	0.994	0.986	0.990
ETTD-CD	0.983	0.995	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.



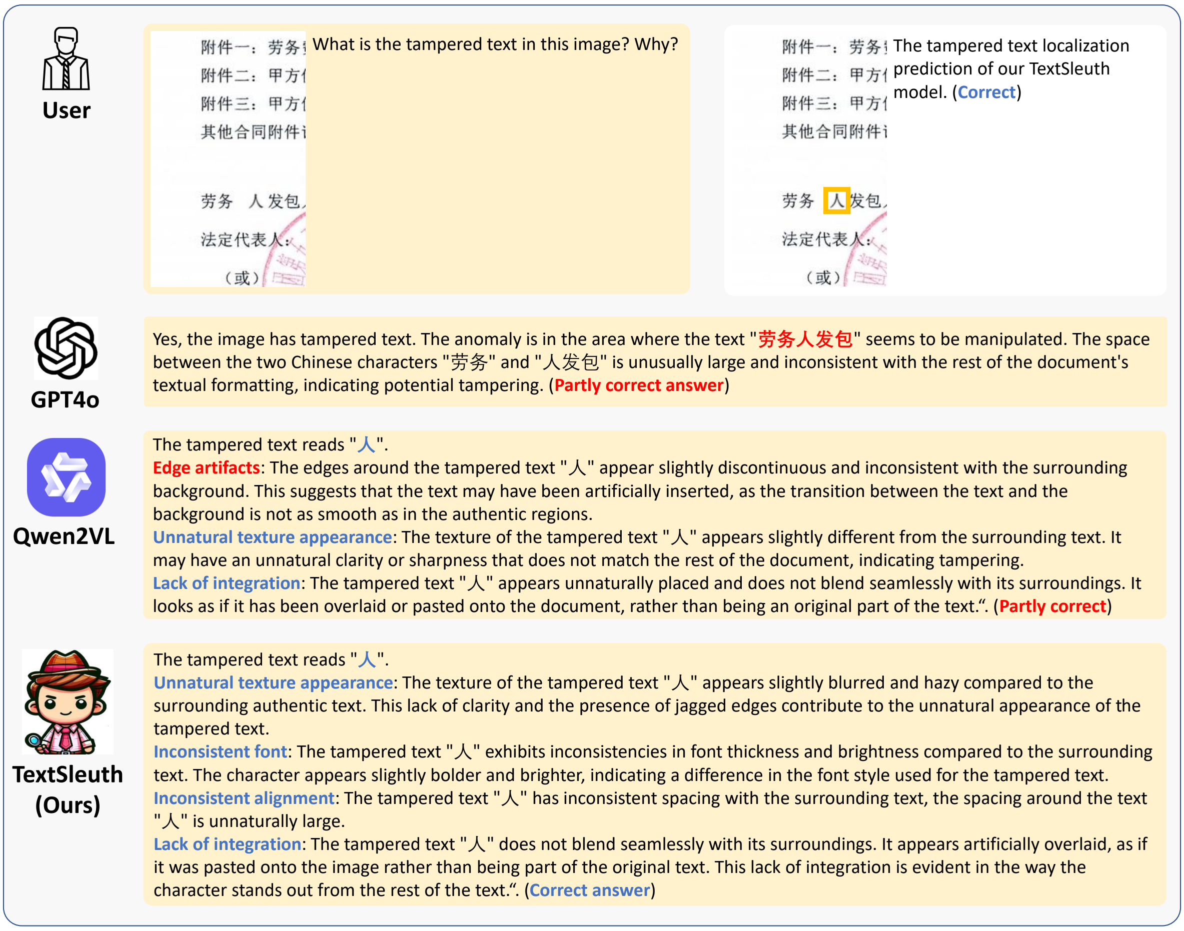
You are an expert good at analyzing tampered text images. You will be provided with two images, **the first is the tampered text image A and the second is the reference image B, with the tampered areas highlighted and the authentic areas darkened**.
Your task is to: First, **recognize the tampered text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

The tampered text reads "GRASS".
Edge Artifacts: The edges around the tampered text "GRASS" appear to be discontinuous with the surrounding background. There is a noticeable boundary around the letters that does not seamlessly blend with the text's immediate environment, suggesting a cut-and-paste effect commonly found in image tampering.
Unnatural Texture Appearance: The texture of the tampered text "GRASS" appears slightly blurred compared to the sharper appearance of the other texts. This distortion effect, where the edges of the letters do not have the same crispness or definition as the authentic text, makes it stand out unnaturally.
Inconsistent Alignment: The spacing between the tampered text "GRASS" and the preceding text "BREAK" appears irregular or has a slight offset when compared to the consistency seen in the other lines of text. This misalignment disrupts the uniformity typically expected in such signage.
Text Incoherence: The text "BREAK GRASS" does not create a coherent or meaningful message within the context of safety instructions, which typically involve actions like "BREAK GLASS." This incoherence adds to the suspicion of tampering, as the revised phrasing results in linguistic nonsense.
Lack of Integration: The tampered text "GRASS" appears as though it has been overlaid or pasted onto the image rather than being a naturally integrated part of the scene. This artificial placement makes the text appear as a foreign element within the image, lacking the integration seen in genuine, unaltered text regions.

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.



User: What is the tampered text in this image? Why?

The tampered text localization prediction of our TextSleuth model. (**Correct**)

GPT4o: Yes, the image has tampered text. The anomaly is in the area where the text "**劳务人发包**" seems to be manipulated. The space between the two Chinese characters "劳务" and "人发包" is unusually large and inconsistent with the rest of the document's textual formatting, indicating potential tampering. (**Partly correct answer**)

Qwen2VL: The tampered text reads "人".

Edge artifacts: The edges around the tampered text "人" appear slightly discontinuous and inconsistent with the surrounding background. This suggests that the text may have been artificially inserted, as the transition between the text and the background is not as smooth as in the authentic regions.

Unnatural texture appearance: The texture of the tampered text "人" appears slightly different from the surrounding text. It may have an unnatural clarity or sharpness that does not match the rest of the document, indicating tampering.

Lack of integration: The tampered text "人" appears unnaturally placed and does not blend seamlessly with its surroundings. It looks as if it has been overlaid or pasted onto the document, rather than being an original part of the text.“. (**Partly correct**)

TextSleuth (Ours): The tampered text reads "人".

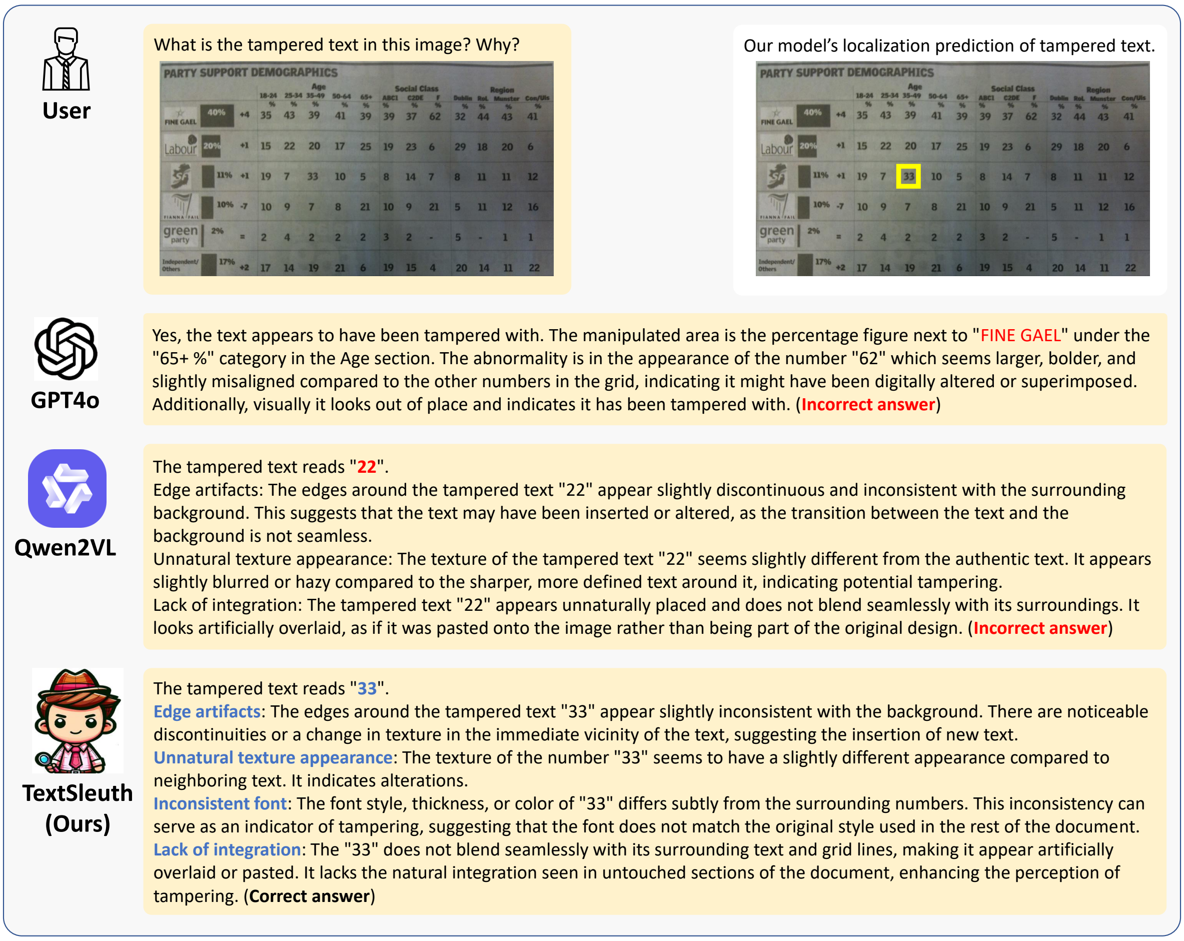
Unnatural texture appearance: The texture of the tampered text "人" appears slightly blurred and hazy compared to the surrounding authentic text. This lack of clarity and the presence of jagged edges contribute to the unnatural appearance of the tampered text.

Inconsistent font: The tampered text "人" exhibits inconsistencies in font thickness and brightness compared to the surrounding text. The character appears slightly bolder and brighter, indicating a difference in the font style used for the tampered text.

Inconsistent alignment: The tampered text "人" has inconsistent spacing with the surrounding text, the spacing around the text "人" is unnaturally large.

Lack of integration: The tampered text "人" does not blend seamlessly with its surroundings. It appears artificially overlaid, as if it was pasted onto the image rather than being part of the original text. This lack of integration is evident in the way the character stands out from the rest of the text.“. (**Correct answer**)

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



What is the tampered text in this image? Why?

PARTY SUPPORT DEMOGRAPHICS

			Age					Social Class			Region			
			18-24 %	25-34 %	35-49 %	50-64 %	65+ %	ABC1 %	C2DE %	F %	Dublin %	RoL %	Munster %	Con/Uls %
FINE GAEL	40%	+4	35	43	39	41	39	39	37	62	32	44	43	41
Labour	20%	+1	15	22	20	17	25	19	23	6	29	18	20	6
SF	11%	+1	19	7	33	10	5	8	14	7	8	11	11	12
FIANNA FAIL	10%	-7	10	9	7	8	21	10	9	21	5	11	12	16
green party	2%	=	2	4	2	2	2	3	2	-	5	-	1	1
Independent/ Others	17%	+2	17	14	19	21	6	19	15	4	20	14	11	22

Our model's localization prediction of tampered text.

PARTY SUPPORT DEMOGRAPHICS

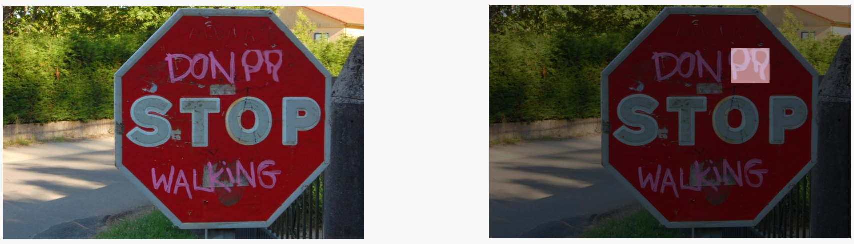
			Age					Social Class			Region			
			18-24 %	25-34 %	35-49 %	50-64 %	65+ %	ABC1 %	C2DE %	F %	Dublin %	RoL %	Munster %	Con/Uls %
FINE GAEL	40%	+4	35	43	39	41	39	39	37	62	32	44	43	41
Labour	20%	+1	15	22	20	17	25	19	23	6	29	18	20	6
SF	11%	+1	19	7	33	10	5	8	14	7	8	11	11	12
FIANNA FAIL	10%	-7	10	9	7	8	21	10	9	21	5	11	12	16
green party	2%	=	2	4	2	2	2	3	2	-	5	-	1	1
Independent/ Others	17%	+2	17	14	19	21	6	19	15	4	20	14	11	22

Yes, the text appears to have been tampered with. The manipulated area is the percentage figure next to "FINE GAEL" under the "65+ %" category in the Age section. The abnormality is in the appearance of the number "62" which seems larger, bolder, and slightly misaligned compared to the other numbers in the grid, indicating it might have been digitally altered or superimposed. Additionally, visually it looks out of place and indicates it has been tampered with. (**Incorrect answer**)

The tampered text reads "**22**".
Edge artifacts: The edges around the tampered text "22" appear slightly discontinuous and inconsistent with the surrounding background. This suggests that the text may have been inserted or altered, as the transition between the text and the background is not seamless.
Unnatural texture appearance: The texture of the tampered text "22" seems slightly different from the authentic text. It appears slightly blurred or hazy compared to the sharper, more defined text around it, indicating potential tampering.
Lack of integration: The tampered text "22" appears unnaturally placed and does not blend seamlessly with its surroundings. It looks artificially overlaid, as if it was pasted onto the image rather than being part of the original design. (**Incorrect answer**)

The tampered text reads "**33**".
Edge artifacts: The edges around the tampered text "33" appear slightly inconsistent with the background. There are noticeable discontinuities or a change in texture in the immediate vicinity of the text, suggesting the insertion of new text.
Unnatural texture appearance: The texture of the number "33" seems to have a slightly different appearance compared to neighboring text. It indicates alterations.
Inconsistent font: The font style, thickness, or color of "33" differs subtly from the surrounding numbers. This inconsistency can serve as an indicator of tampering, suggesting that the font does not match the original style used in the rest of the document.
Lack of integration: The "33" does not blend seamlessly with its surrounding text and grid lines, making it appear artificially overlaid or pasted. It lacks the natural integration seen in untouched sections of the document, enhancing the perception of tampering. (**Correct answer**)

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.

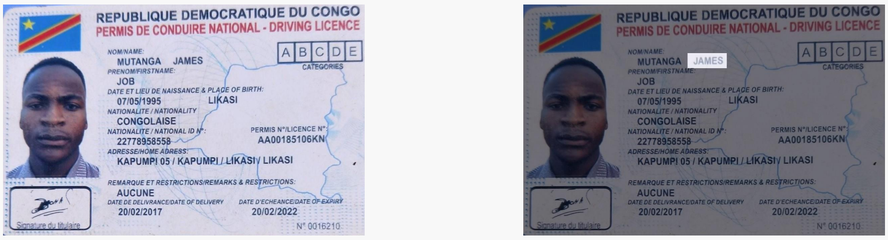
Unnatural Texture Appearance: The texture of the tampered "PR" text appears somewhat hazy and less sharp than the surrounding authentic texts. This indicates manipulation, as the integrated texture should normally appear consistent.

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.



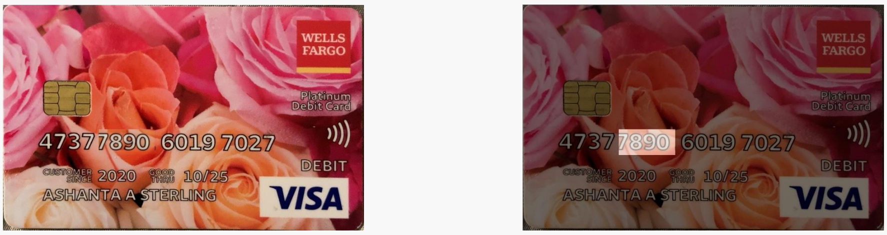
The tampered text reads "JAMES".

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.



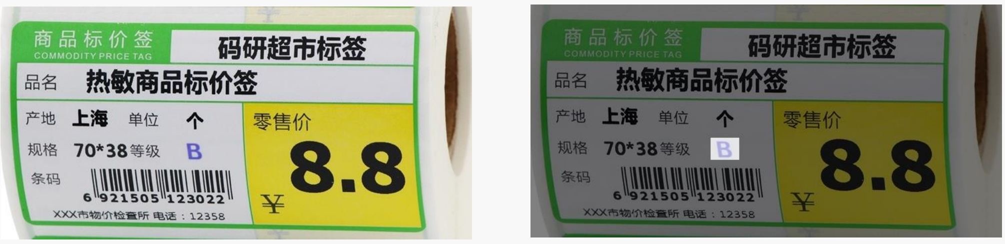
The tampered text reads "7890".

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.

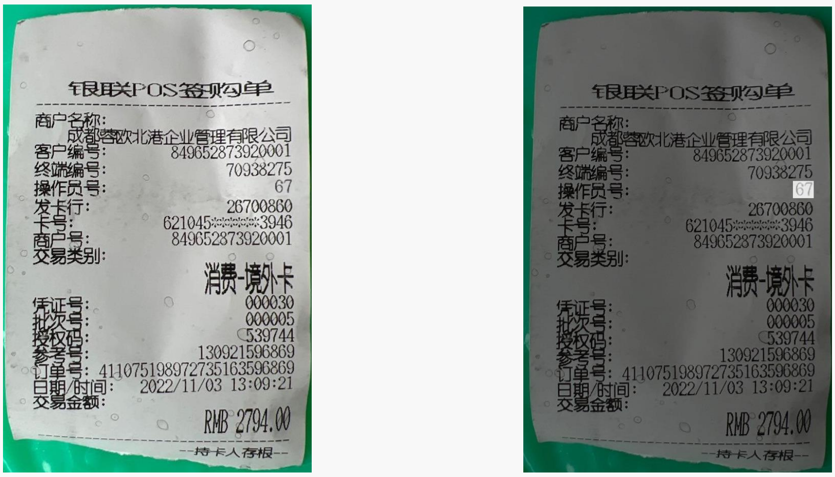


商品标价签
COMMODITY PRICE TAG
码研超市标签
品名 热敏商品标价签
产地 上海 单位 个
规格 70*38 等级 B
条码 6 921505 123022
零售价
¥ 8.8
XXX市物价检查所 电话：12358

The tapered text reads "B".

Unnatural Texture Appearance: The letter "B" displays an unnatural texture. This is evident through slight blurriness, jagged edges, or a distortion effect making it stand out against the surrounding authentic text.

Inconsistent Font: The character "B" differs in font size, thickness, brightness, and color compared to the surrounding text. This variance makes it appear distinct and separate from the rest of the text elements in the label.



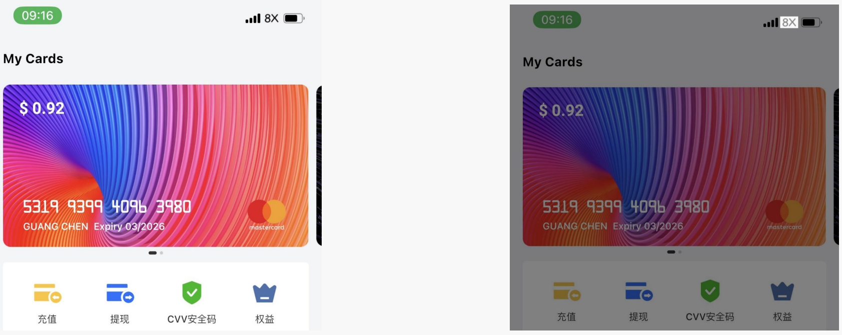
银联POS签购单
商户名称:
成都蓉欧北港企业管理有限公司
客户编号: 849652873920001
终端编号: 70938275
操作员号: 67
发卡行: 26700860
卡号: 621045******3946
商户号: 849652873920001
交易类别:
消费-境外卡
凭证号: 000030
批次号: 000005
授权码: 539744
参考号: 130921596869
订单号: 411075198972735163596869
日期/时间: 2022/11/03 13:09:21
交易金额:
RMB 2794.00
--持卡人存根--

The tampered text reads "67".

Edge Artifacts: The number "67" has visible edge artifacts, indicating tampering. The boundaries around these characters appear more prominent or sharper than the surrounding text, suggesting they were added separately.

Inconsistent Font: The font used for "67" shows slight variances in thickness or brightness compared to the authentic text, indicating that it was not originally part of the document.

Unnatural Texture Appearance: The "67" appears slightly blurred or hazy compared to nearby numbers. This effect might result from editing that failed to seamlessly integrate the new text with the texture of the paper.



09:16
8X
My Cards
$ 0.92
5319 9399 4096 3980
GUANG CHEN Expiry 03/2026
mastercard
充值
提现
CVV安全码
权益

The tampered text reads "8X".

Edge Artifacts: The edges around the "8X" tampered text appear slightly inconsistent with the smooth and continuous background typically seen in authentic regions. This manifests in subtle edge discontinuities or abrupt transitions that do not blend seamlessly with the surrounding area.

Unnatural Texture Appearance: The "8X" text shows signs of unnatural texture compared to its original appearance. This includes potential blurriness, haziness, or jagged contours which are not naturally found on the finely rendered elements typical of this user interface.

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.