Detecting Text Manipulation in Images using Vision Language Models

Vidit Vidit
vidit.vidit@idiap.ch
Pavel Korshunov
pavel.korshunov@idiap.ch
Amir Mohammadi
amir.mohammadi@idiap.ch
Christophe Ecabert
christophe.ecabert@idiap.ch
Ketan Kotwal
ketan.kotwal@idiap.ch
Sébastien Marcel
sebastien.marcel@idiap.ch

IDIAP Research Institute
Martigny
Switzerland

Abstract

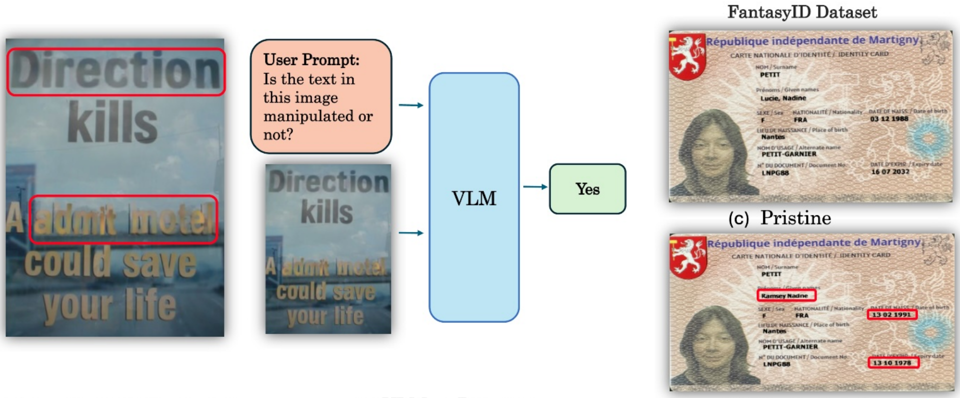
Recent works have shown the effectiveness of Large Vision Language Models (VLMs or LVLMs) in image manipulation detection. However, text manipulation detection is largely missing in these studies. We bridge this knowledge gap by analyzing closed- and open-source VLMs on different text manipulation datasets. Our results suggest that open-source models are getting closer, but still behind closed-source ones like GPT-4o. Additionally, we benchmark image manipulation detection-specific VLMs for text manipulation detection and show that they suffer from the generalization problem. We benchmark VLMs for manipulations done on in-the-wild scene texts and on fantasy ID cards, where the latter mimic a challenging real-world misuse. Paper Page: https://www.idiap.ch/paper/textvlmdet/

1 Introduction

Advancements in image generation has made it easier to edit and generate realistic looking images. However, this capability comes with a major downside, the same tools can be exploited for malicious purposes, such as generating fake and manipulated images. Such images are increasingly being used to spread misinformation [8] or create fraudulent identities that bypass online Know Your Customer (KYC) checks[1]. Typically, these manipulations involve addition/removal of objects in an image through a generative method and can be followed by post-processing to seamlessly blend the altered regions. With rapid progress in generation quality, these manipulated images have become more difficult to visually identify,

[1] https://securityexpress.info/fake-passport-generated-by-chatgpt-bypasses-security/

especially, when small but semantically critical regions, such as text, are modified. Detecting such subtle changes is challenging and current image forgery detection methods often overlook the manipulated text regions. One of the goals of this work is to bridge this gap.


FantasyID Dataset
Direction kills A admit motel could save your life
User Prompt: Is the text in this image manipulated or not?
VLM
Yes
République indépendante de Martigny
CARTE NATIONALE D'IDENTITÉ / IDENTITY CARD
PETIT
Lucie, Nadine
Ramsey Nadine
03 12 1988
13 02 1991
16 07 2032
13 10 1978
PETIT-GARNIER
Nantes
LNPG88

(c) Pristine

(a) Text in Wild Manipulation (b) VLM as Detector (d) Manipulated

Fig 1: **VLM for Text Manipulation Detection:** (a) Example from OSTF [26] dataset of in-the-wild text manipulation (in red). (b) With the help of a user prompt, we ask a pretrained VLM whether the accompanying image contains a text manipulation. The output of VLM is then used as a label for binary classification. (c,d) Example from FantasyID [15] dataset which simulate the real-world scenario of text manipulation in ID documents. (d) The altered text is shown in red. Best viewed digitally.

Recent works have shown promising results in the detection of generated and manipulated images [3, 10, 16, 18, 19, 27, 32, 33]. Typically, trained models can provide an image-level label and/or localize tampered regions within the image. With the growing adoption of large vision language models (VLMs), they started being used to detect generated or tampered images [12, 14, 31]. One useful aspect of such models is their zero-shot reasoning ability, as well as their capacity to generate textual explanations for their prediction. VLM-based methods have shown significant improvement over smaller convolutional or transformer-based manipulation detection models for face presentation attack detection [24].

As VLMs are being employed for a variety of tasks, it becomes pertinent to understand where they stand on text manipulation detection as well. In this work, we would like to answer few questions like *how big is the gap between open and closed source models?* and *how well do the finetuned VLMs generalize to text manipulation detections?*. We also illustrate how the prompt design and input image resolution affect the final result. These takeaways can help in designing a better text manipulation detectors.

Although there are existing datasets with text manipulation of images taken in the wild (OSTF [26]), the main focus of this work is on detecting the real-world text forgery in ID documents, since it is a critical security risk in the prevalence of online-based enrollment and KYC checks. For this purpose, evaluate how practical are current VLMs for detecting tampered text in ID documents, using FantasyID [15]. This data set consists of the manipulation of text fields on *fantasy* id cards, which simulates an *injection-attack* by a fraudster.

To summarize our contributions, (a) we study open and closed source VLMs for text manipulation detection task in zero-shot setting, (b) we evaluated existing image manipulation specific finetuned VLMs on their generalization capability to text manipulation detection,

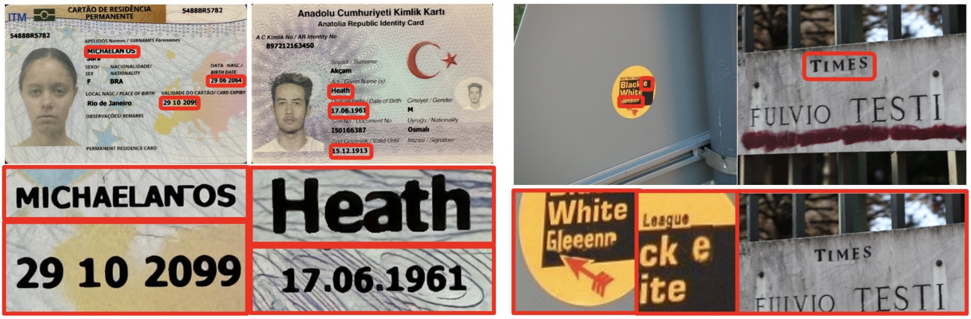


Fig 2: **FantasyID [15] and OSTF [26]** Two different templates from our proposed, FantasyID (left) and OSTF (right) dataset. The red box show the manipulated regions and their corresponding zoomed-in image. (Left) Altered text mostly show jagged edges due to inconsistent blending with the background. (Right) Zoomed-in view of the tampered region where some of the manipulation artifacts are visible. Such regions occupy small area of the full image making it a challenging detection task.

and (c) to promote research in real-world forgery detection, we benchmark different VLMs on FantasyID [15] and OSTF [26]. We will make our code public upon acceptance.

2 Related Work

Image Manipulation Datasets. Several training and benchmark datasets have been proposed for image manipulation detection. Common datasets include Photoshop-edited and deepfake collections such as CASIAv2 [7], FantasticReality [13], Columbia [22], COVERAGE [30], IMD2020 [23], and FaceApp [5]. These datasets offer limited manipulation diversity and are primarily used for evaluation rather than large-scale training. FakeShield [31] and SIDA [12] introduce large-scale AI-generated image datasets (based on GANs or diffusion models) for training. However, they lack sufficient evaluation of text-based manipulations. Open-Set Text Forensics (OSTF) [26] addresses this by introducing a dataset with text edits generated using both CNN and diffusion models. We also study ID card datasets to better understand detection performance on the real-world document tampering. PAD [28] propose ID card datasets for presentation attack detection scenarios. In contrast, we target digital manipulation reflecting an injection attack scenario. While PAD [28] involves manipulation of original government IDs, which is legally prohibited, FantasyID [15] avoids legal issues by creating fantasy templates and using images with appropriate licenses. In this work, we evaluate detection methods on both the OSTF dataset and FantasyID.

Image Manipulation Detection. Various works [3, 10, 16, 18, 19, 27, 32, 33] have proposed detection methods targeting different types of forgeries, such as *inpainting*, *copy-move*, and *splicing*. Detection tasks can involve image-level classification or localization of tampered regions. Typically, training datasets contain pristine and manipulated images, with or without annotations indicating the manipulated areas. These manipulations are introduced using CNN-based generative methods or diffusion-based approaches. TruFor [10] learns camera-specific noise patterns and identifies tampered regions by detecting inconsistencies in these patterns. HiFi-Net [11] detects image alterations using multi-branch feature extraction. MVSS-Net [6] utilizes multi-scale, multi-branch features to better capture image noise, while IML-ViT [21] employs a transformer-based architecture with edge-guided loss to enhance focus on tampering artifacts. EditGuard [34] embeds watermarks into images and later verifies authenticity based on their presence. These methods are trained on specific types of tampering and aim to generalize to unseen manipulation techniques. However, as generative models improve, generalizing to novel manipulations becomes increasingly challenging. Text tampering detection methods [25, 26, 29] highlight the importance of benchmarks tailored to text manipulation, as such edits often affect small image regions. Nevertheless, similar to prior approaches, these methods also struggle to generalize across new editing techniques.

Vision-Language Models for Detection. The use of large Vision-Language Models (VLMs) has grown rapidly, largely due to their zero-shot capabilities. Trained on large-scale internet datasets, these models can quickly adapt to downstream applications. While proprietary models such as GPT-4o[2] have demonstrated strong performance on a range of benchmarks [4], open-source alternatives like Qwen-VL [1] and Llama-Vision [9] are closing the gap. These transformer-based models tokenize images into patch sequences and project them into a shared embedding space with text features, enabling rich cross-modal representations. Recently, VLMs have been applied to detect and localize image manipulations with FakeShield [31] and SIDA [12], fine-tuned LLaVA [20] models, not only achieving state-of-the-art performance but also generating textual justifications for their predictions, enhancing interpretability. Here, we study these VLMs for text manipulation detection which is largely missing in their evaluations.

3 Experimental Setup

We evaluate GPT-4o which is a closed source VLM, Qwen-VL [1] and Llama-Vision-3.2 [9] which are generic open source VLM, through zero-shot prompting task. As shown in Fig. 1, we prompt query image with a detailed prompt asking VLM to generate a final label. We further study the finetuned VLMs, FakeShield [31] and SIDA [12], on their generalization to text manipulation detection. We use their default prompts to get the prediction of the query image. Additionally, we compare with a non-VLM baseline TruFor [10], because it has a good generalization capability to unseen tampering methods.

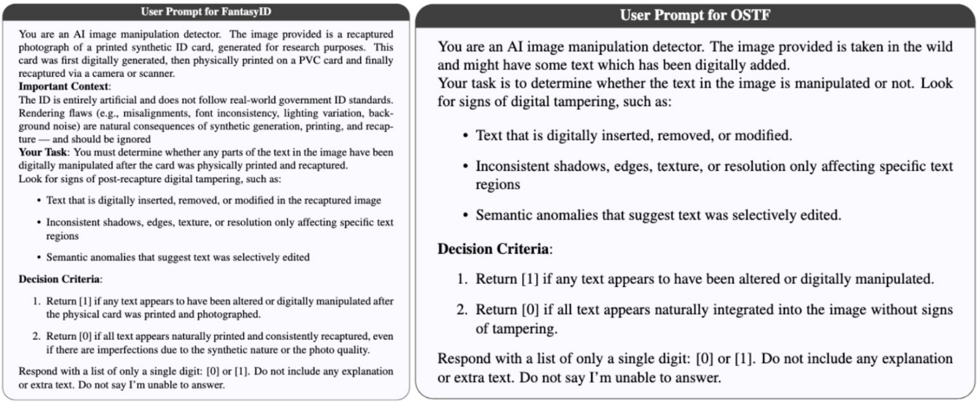
3.1 Datasets

Most VLM-based manipulation detection methods lack a study on text specific manipulations. This task is challenging because usually text regions occupy smaller area in the image (Fig. 2). Therefore, we present their zero-shot evaluation on the following datasets:

Open-Set Text Forensics (OSTF) [26] dataset consists of images with in-the-wild text manipulation, which are sourced from 4 different scene text datasets, and manipulation is done using 8 different methods comprising text rendering, a convolution, and diffusion model-based editing. We use their test split with 2238 images (1267 pristine, 971 manipulated) for our evaluation.

FantasyID [15] Text manipulation in sensitive documents poses severe security risks for KYC applications. This dataset has 362 ID cards which mimic real-world IDs. IDs are printed on PVC cards and captured using 3 camera devices. Unlike OSTF [26], the text-manipulated images by fine-tuning TextDiffuser-2 [2] on pristine cards. This creates difficult

[2] https://platform.openai.com/docs/models/gpt-4o



User Prompt for FantasyID

You are an AI image manipulation detector. The image provided is a recaptured photograph of a printed synthetic ID card, generated for research purposes. This card was first digitally generated, then physically printed on a PVC card and finally recaptured via a camera or scanner.
Important Context:
The ID is entirely artificial and does not follow real-world government ID standards. Rendering flaws (e.g., misalignments, font inconsistency, lighting variation, background noise) are natural consequences of synthetic generation, printing, and recapture — and should be ignored
Your Task: You must determine whether any parts of the text in the image have been digitally manipulated after the card was physically printed and recaptured.
Look for signs of post-recapture digital tampering, such as:

- Text that is digitally inserted, removed, or modified in the recaptured image
- Inconsistent shadows, edges, texture, or resolution only affecting specific text regions
- Semantic anomalies that suggest text was selectively edited

Decision Criteria:

1. Return [1] if any text appears to have been altered or digitally manipulated after the physical card was printed and photographed.
2. Return [0] if all text appears naturally printed and consistently recaptured, even if there are imperfections due to the synthetic nature or the photo quality.

Respond with a list of only a single digit: [0] or [1]. Do not include any explanation or extra text. Do not say I'm unable to answer.

User Prompt for OSTF

You are an AI image manipulation detector. The image provided is taken in the wild and might have some text which has been digitally added.
Your task is to determine whether the text in the image is manipulated or not. Look for signs of digital tampering, such as:

- Text that is digitally inserted, removed, or modified.
- Inconsistent shadows, edges, texture, or resolution only affecting specific text regions
- Semantic anomalies that suggest text was selectively edited.

Decision Criteria:

1. Return [1] if any text appears to have been altered or digitally manipulated.
2. Return [0] if all text appears naturally integrated into the image without signs of tampering.

Respond with a list of only a single digit: [0] or [1]. Do not include any explanation or extra text. Do not say I'm unable to answer.

Fig 3: **VLM Prompt for FantasyID [15] and OSTF [26]** We provide the necessary context of the problem along with intended output format.

to detect manipulation as the artifacts (Fig. 2) are subtle. For our evaluations, we use the 1572 images (786 pristine, 786 manipulated).

3.2 Baselines

TruFor [10] uses a multi-branch Transformer encoder architecture to combine features from RGB images and Noiseprint++ images to predict an anomaly localization map, a confidence map, and a final score. Noiseprint++ [10] is a fully convolutional network trained to extract subtle noise in pristine images caused by imperfections in camera hardware or in-camera processing steps. When an image is manipulated, it creates a different noise pattern in the manipulated region, which can be treated as an anomaly. This is used to predict a final score for the image. Trufor can accept images of arbitary sizes but we consider two variants (a) TruFor-*low* longest side is resized to 512 (b) TruFor-*high*, the longest side is resized to 768 only for the images that cannot be processed in the original image dimension.
SIDA [12] is based on LISA [17] model, which the authors fine-tuned on a large dataset of manipulated images collected from social media. The model generates an explanation of manipulation along with its mask. We use the label provided by the model for evaluation. We use the default prompt with the query image. For input, the images are split into 4 non-overlapping patches of size 336^2 and then processed by vision encoder.
FakeShield [31] proposes a fine-tuned LLaVA-v1.5-13B [20] model for manipulation detection and localization, which is tuned on multimodal dataset of manipulated images to generate description of manipulated region along with its localization mask. The training data consists of edited images based on deepfakes, photoshop, and diffusion models. In our evaluations, we use the default prompt and consider an image manipulated when the model generates a mask. The input image size is the same as in SIDA.
Llama-Vision We benchmark the open-source Llama-3.2-90B model with vision capability in the zero-shot setting. It is trained on general image reasoning tasks with webscale data, and we prompt this model to behave as a text manipulation detector with temperature set to 0 for *deterministic* answer. The input image are split into up to 4 square patches of size 560.

Qwen-VL [1] is another powerful open-source general VLM. Similarly to Llama, we prompt this model (Qwen2.5-VL-72B version) to behave as a text manipulation detector in the zero-shot setting. While we keep the default image resolution, we set the model temperature to 0 to provide a *deterministic* answer. Qwen model can take images of arbitrary sizes and process them as a sequence of 28^2 non-overlapping patches.

GPT-4o We evaluate the closed-source GPT-4o model (`gpt-4o-2024-11-20` version), for the manipulation detection task. OpenAI's API allows a query with images in two resolutions: *low* and *high*. The former process image resolution with the longest side of 512, while the latter can handle 768(short-side) × 2000(long-side). The images are dynamically resized or padded to fit these dimensions while maintaining the aspect ratio. The model temperature is set to 0 to provide a *deterministic* answer.

3.3 Metrics and compute details

We measure binary classification performance using F_1 scores and Avg F_1. The confidence threshold of 0.5 is set for TruFor baselines. We do not measure localization accuracy as it is non-trivial to convert the output from GPT-4o, Qwen and Llama into a mask. This mask generation is beyond the scope of the work. For open-source models, we use a single H100 GPU, and GPT-4o was accessed through the OpenAI API, whose total experiment cost was around $200.

3.4 Prompt Design

For GPT-4o, Llama-Vision, and Qwen-VL, we take a zero-shot approach in which, given a prompt and a query image, VLM provides a final prediction for the image. This prediction is then used as a label in binary classification. We need to design an appropriate prompt that describes the underlying task and the required output format. The prompts used for the experiment are shown in Fig. 3. FantasyID prompts are more detailed than OSTF as they are *fantasy* IDs and can be considered as fake, if not explicitly prompted to ignore that fact. The prompts for FakeShield [31] and SIDA [12] are the same as proposed by the authors. Since these are specialized for manipulation detection task, they do not need different prompts.

4 Results

Text Manipulation Detection. We evaluate our baselines with the designed prompts and Tab. 1 report the performance on the two datasets showing GPT-4o to be the best performing model overall.

- *closed vs open source VLM:* Both Qwen-VL and Llama-3.2-Vision models underperform compared to GPT-4o. Avg F_1 for GPT-4o is higher than Qwen-VL by 29% on FantasyID and 7% on OSTF dataset Tab. 1. Despite having access to the original image resolution, Qwen-VL still underperforms even compared to GPT-4o-low version, which often downsizes input images. Among open-source models, Qwen-VL performs consistently better than Llama-3.2-Vision, similar to other reported benchmarks [4]. Clearly, the open source models are lagging behind ChatGPT but Qwen-VL can be a competitive choice.

- *general vs specialized VLM:* FakeShield [31] and SIDA [12] underperform compared to all the general VLMs: Llama, Qwen and GPT-4o. The reason could be that specialized VLMs are finetuned versions of a smaller LLaVa-13B [20] model. FakeShield

Model	OSTF			FantasyID		
	$F_1(P)$	$F_1(M)$	Avg F_1	$F_1(P)$	$F_1(M)$	Avg F_1
TruFor-low [10]	0.72	0.17	0.45	0.64	0.20	0.42
TruFor-high [10]	0.74	0.56	0.65	0.70	0.72	0.71
FakeShield [31]	0.51	0.51	0.51	0.51	0.51	0.51
SIDA [12]	0.70	0.24	0.47	0.67	0.01	0.34
Llama-3.2-90b-Vision	0.75	0.52	0.64	0.65	0.27	0.46
Qwen-2.5-VL-72b [1]	0.85	0.74	0.79	0.72	0.40	0.56
GPT-4o-low	0.87	0.82	0.84	0.74	0.50	0.62
GPT-4o-high	0.86	0.85	**0.86**	0.84	0.86	**0.85**

Tab 1: **Evaluation on OSTF [26] and FantasyID [15]:** We compare the performance of different VLMs and a non-VLM baseline TruFor using Avg.F_1. Note: $F_1(P)$: F_1 for Pristine class, $F_1(M)$: F_1 for Manipulated class

and SIDA are limited by both the reasoning capacity of the smaller LLM and the vision encoder, which can only handle smaller image resolution. However, they had competitive results on their own scene/object manipulation benchmarks [12, 31], which also suggests that they suffer from the generalization problem. In their training and evaluation datasets, they rarely see text manipulation, making them biased towards artifacts of the scene manipulations.

- *VLM vs TruFor:* TruFor [10] is a non-VLM baseline with a good out of distribution generalization. Here we see that compared to GPT-4o, it underperforms on both low and high versions. However, it is better than FakeShield [31] and SIDA [12] on OSTF but similar on FantasyID. It shows that TruFor, while being a smaller model, can be competitive when compared to specialized VLMs. It is worse than Qwen-VL and Llama-Vision on OSTF but better on FantasyID when using higher image resolution. We would like to emphasize that TruFor is trained to detect manipulation, whereas Qwen-VL and Llama-Vision are zero-shot prompted. While being competitive with 'high' resolution, TruFor suffers under 'low' image resolution. This results in a random accuracy for FantasyID with tiny regions of manipulated text, demonstrating the challenges of manipulation detection.

In summary, closed VLM GPT-4o shows significantly higher accuracy on both datasets compared to both open and specialized VLMs. Qwen-VL remains a competitive open source option, but text manipulation detection remains a hard problem for specialized VLMs and TruFor to generalize. However, TruFor performs better than specialized VLMs on both datasets. For fair comparison, we do not include OSTF [26] as it is a purely localization baseline.

Prompts Matter. The behavior of VLM is usually controlled by its prompts. This becomes crucial when we are evaluating its performance in zero-shot setting. We ablate the influence of prompts(temperature set as 0) on GPT-4o responses . To this end, we design prompts with varying levels of description about the text manipulation task.

- L1: this prompt only mentions that the query image may consist of some digital manipulations, without describing the nature of manipulation. The prompt is as follows:

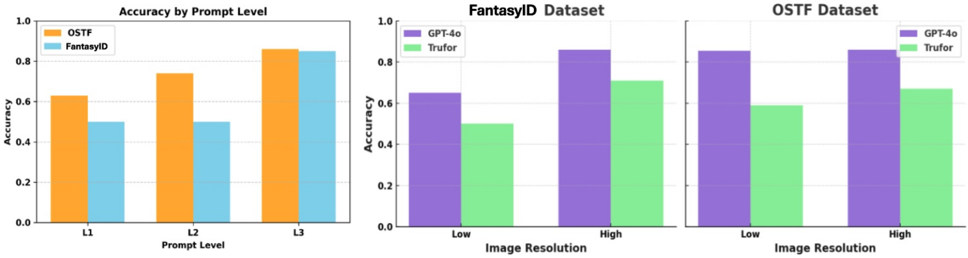


Fig 4: **Effect of Descriptive Prompts and Image Resolution:** (Left) With more detailed prompts, we can capture the task context well. (Right) Both GPT-4o and TruFor get a boost in their performance when image resolution is high. High resolution helps them capture smaller artifacts well.

> You are an AI image manipulation detector. The image provided is taken in the wild and might have some digital manipulations. Your task is to determine whether the image is manipulated or not?

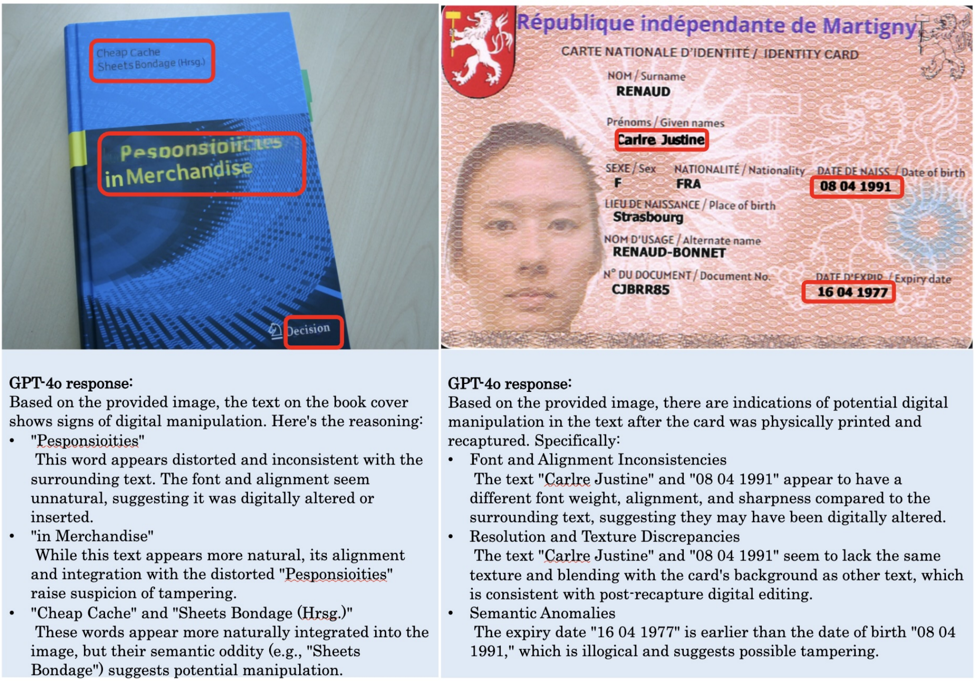
- L2: Here, we mention that the text in the image can be manipulated and clearly mention the task is to determine whether the text regions are altered. The prompt is the following:

> You are an AI image manipulation detector. The image provided is taken in the wild and might have text digitally altered. Your task is to determine whether any of the text in the image is manipulated or not?

- L3: Lastly, we give a detailed description on the text manipulation task and add a few possible artifacts that can be attributed to the alteration process. This is exactly the prompt as in Fig. 3.

Fig. 4 summarizes the influence of prompts on GPT-4o prediction capability. The accuracy of GPT-4o model increases as the prompts get more detailed. For OSTF [26] dataset, there is a consistent increase in the performance as prompts get detailed. However, for FantasyID, both L1 and L2 fail to capture the *fantasy* nature of the ID cards and GPT-4o labels all the images as manipulated. This shows that performance can widely vary if prompts do not describe the underlying context of the task well.

Importance of Image Resolution. Since text regions often occupy a relatively small area of the image, it is crucial that the model possesses sufficient representational power to capture such fine details. As discussed in Sec. 3.2, the input image resolution varies across different models, which can significantly impact the performance on text manipulation detection tasks. Both FakeShield [31] and SIDA [12] operate on lower-resolution image patches compared to Llama and Qwen models, limiting their ability to capture subtle artifacts. This limitation correlates with their lower performance, as shown in Tab. 1. To further illustrate this observation, we analyze the performance of the GPT-4o and TruFor [10] models by explicitly varying the input image resolution (see Sec. 3.2 for details). As shown in Fig. 4, performance on both datasets improves when higher-resolution images are used. This effect is particularly significant in the case of the FantasyID dataset, where manipulation artifacts become indiscernible at lower resolutions, leading to random performance from TruFor [10].


Cheap Cache
Sheets Bondage (Hrsg.)
Pesponsioiities
in Merchandise
Decision
République indépendante de Martigny
CARTE NATIONALE D'IDENTITÉ / IDENTITY CARD
NOM / Surname
RENAUD
Prénoms / Given names
Carlre Justine
SEXE / Sex
F
NATIONALITÉ / Nationality
FRA
DATE DE NAISS. / Date of birth
08 04 1991
LIEU DE NAISSANCE / Place of birth
Strasbourg
NOM D'USAGE / Alternate name
RENAUD-BONNET
N° DU DOCUMENT / Document No.
CJBRR85
DATE D'EXPIR. / Expiry date
16 04 1977

GPT-4o response:
Based on the provided image, the text on the book cover shows signs of digital manipulation. Here's the reasoning:

- "Pesponsioities"
  This word appears distorted and inconsistent with the surrounding text. The font and alignment seem unnatural, suggesting it was digitally altered or inserted.
- "in Merchandise"
  While this text appears more natural, its alignment and integration with the distorted "Pesponsioities" raise suspicion of tampering.
- "Cheap Cache" and "Sheets Bondage (Hrsg.)"
  These words appear more naturally integrated into the image, but their semantic oddity (e.g., "Sheets Bondage") suggests potential manipulation.

GPT-4o response:
Based on the provided image, there are indications of potential digital manipulation in the text after the card was physically printed and recaptured. Specifically:

- Font and Alignment Inconsistencies
  The text "Carlre Justine" and "08 04 1991" appear to have a different font weight, alignment, and sharpness compared to the surrounding text, suggesting they may have been digitally altered.
- Resolution and Texture Discrepancies
  The text "Carlre Justine" and "08 04 1991" seem to lack the same texture and blending with the card's background as other text, which is consistent with post-recapture digital editing.
- Semantic Anomalies
  The expiry date "16 04 1977" is earlier than the date of birth "08 04 1991," which is illogical and suggests possible tampering.

Fig 5: **GPT-4o response on OSTF and FantasyID:** We prompt GPT-4o to give a detailed response for its reasoning on text region manipulation. It is able to capture the manipulated regions(shown in red) based on inconsistency with the neighboring regions and other texts in the image. (Left) However, it does *fail* to retrieve all the manipulated texts, bottom right text "Decision" is missing in its reasoning. (Right) Beyond texture and other pixel-wise artifacts, it can also detect semantic errors, as mentioned the expiry date is before birthdate.

GPT-4o explanations. GPT-4o performs much better compared to other the models. This is not surprising since it is the best performing VLM [4] on diverse benchmarks. To better understand reason for its performance, we prompt GPT-4o to follow Fig. 3 and ask it to answer in detail. The corresponding response is shown in Fig. 5. For both datasets, it generates an accurate description of textual tampering by analyzing variation in texture patterns, text alignment and semantic coherence in the image. Although, it identifies manipulated images well, it can still miss some manipulated regions. This issue can be better handled in the prompt by asking GPT-4o to single out all the manipulated words firsts and then individual verify for manipulation artifacts.

5 Conclusion

In this paper, we highlight the motivation and challenges of text manipulation detection. With our experiments, we bridge the gap of lack of text manipulation evaluations in the current literature. As VLMs are getting popular for image manipulation detection, we evaluate several models and can conclude the following: (a) The gap between closed source, GPT-4o and open source VLMs, Qwen-VL, Llama-Vision is large. The best performing open source model Qwen-VL trails behind GPT-4o by 7% on OSTF [26] and 29% on FantasyID.

To reduce this gap, the responses by GPT-4o can be used to distill the knowledge into open source models and improve them. (b) Specialized VLMs FakeShield [31] and SIDA [12] fail to generalize to text manipulation. This highlights the challenges of text manipulation and the need of diverse training/evaluation datasets. (c) VLM as a text manipulation detector works well when prompted with the right context and when they can process images in higher resolution. The underlying vision encoder and LLM affect the final performance.(d) FantasyID evaluation show that document manipulation detection is a challenging task, and most VLMs underperform.

We hope that research community will benefit from our findings and the proposed new dataset to improve VLMs for manipulation detection tasks.

6 Acknowledgment

This work was funded by InnoSuisse 106.729 IP-ICT ROSALIND project.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024.

[3] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021.

[4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[6] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022.

[7] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pages 422–426. IEEE, 2013.

[8] Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, et al. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *arXiv preprint arXiv:2405.11697*, 1(8), 2024.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023.

[11] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[12] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. *arXiv preprint arXiv:2412.04292*, 2024.

[13] Vladimir V Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. *Advances in neural information processing systems*, 32, 2019.

[14] Alain Komaty, Hatef Otroshi Shahreza, Anjith George, and Sebastien Marcel. Exploring chatgpt for face presentation attack detection in zero and few-shot in-context learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[15] Pavel Korshunov, Amir Mohammadi, Vidit Vidit, Christophe Ecabert, and Sébastien Marcel. Fantasyid: A dataset for detecting digital manipulations of id-documents, 2025. URL https://arxiv.org/abs/2507.20808.

[16] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 375–384, 2021.

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[18] Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 8301–8310, 2019.

[19] Yuanman Li and Jiantao Zhou. Fast and effective image copy-move forgery detection via hierarchical feature point matching. *IEEE Transactions on Information Forensics and Security*, 14(5): 1307–1322, 2018.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[21] Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

[22] Tian-Tsong Ng, Jessie Hsu, and Shih-Fu Chang. Columbia image splicing detection evaluation dataset. *DVMM lab. Columbia Univ CalPhotos Digit Libr*, 2009.

[23] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pages 71–80, 2020.

[24] Guray Ozgur, Eduarda Caldeira, Tahar Chettaoui, Fadi Boutros, Raghavendra Ramachandra, and Naser Damer. Foundpad: Foundation models reloaded for face presentation attack detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 745–755, February 2025.

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

[26] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 694–702, 2025.

[27] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018.

[28] Juan E Tapia, Naser Damer, Christoph Busch, Juan M Espin, Javier Barrachina, Alvaro S Rocamora, Krištof Ocvirk, Leon Alessio, Borut Batagelj, Sushrut Patwardhan, et al. First competition on presentation attack detection on id card. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024.

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

[30] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016.

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

[32] Qichao Ying, Hang Zhou, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Learning to immunize images for tamper localization and self-recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13814–13830, 2023.

[33] Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. Cross: Diffusion model makes controllable, robust and secure image steganography. *Advances in Neural Information Processing Systems*, 36:80730–80743, 2023.

[34] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11964–11974, 2024.