CheckGuard: Advancing Stolen Check Detection with a Cross-Modal Image-Text Benchmark Dataset

Fei Zhao*
The University of Alabama at Birmingham
Birmingham, USA
larry5@uab.edu

Jiawen Chen*
Beijing-Dublin International College, Beijing University of Technology
Beijing, China
cnjywr@emails.bjut.edu.cn

Bin Huang
Ji Zhi Xing Huo Technology Beijing
Beijing, China
huangbin.cn@gmail.com

Chengcui Zhang
The University of Alabama at Birmingham
Birmingham, USA
czhang02@uab.edu

Gary Warner
The University of Alabama at Birmingham
Birmingham, USA
gar@uab.edu

Abstract

The prevalence of check fraud, particularly with stolen checks sold on platforms such as Telegram, creates significant challenges for both individuals and financial institutions. This underscores the urgent need for innovative solutions to detecting and preventing such fraud on social media platforms. While deep learning techniques show great promise in detecting objects and extracting information from images, their effectiveness in addressing check fraud is hindered by the lack of comprehensive, open-source, large training datasets specifically for check information extraction. To bridge this gap, this paper introduces "CheckGuard," a large labeled image-to-text cross-modal dataset designed for check information extraction. CheckGuard comprises over 7,000 real-world stolen check image segments from more than 15 financial institutions, featuring a variety of check styles and layouts. These segments have been manually labeled, resulting in over 50,000 samples across seven key elements: Drawer, Payee, Amount, Date, Drawee, Routing Number, and Check Number. This dataset supports various tasks such as visual question answering (VQA) on checks and check image captioning. Our paper details the rigorous data collecting, cleaning, and annotation processes that make CheckGuard a valuable resource for researchers in check fraud detection, machine learning, and multimodal large language models (MLLMs). We not only benchmark state-of-the-art (SOTA) methods on this dataset to assess their performance but also explore potential enhancements. Our application of parameter-efficient fine-tuning (PEFT) techniques on the SOTA MLLMs demonstrates significant performance improvements, providing valuable insights and practical approaches for enhancing model efficacy on this task. As an evolving project, CheckGuard will continue to be updated with new data, enhancing its utility and driving further advancements in the field. Our PEFT-based MLLM code is available at: https://github.com/feizhao19/CheckGuard. For data access, researchers are required to contact the authors directly.

CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Information systems** → **Multimedia and multimodal retrieval**.

Keywords

Machine Learning, Multi-modal Large Language Model, Cross-modal Generation, Check Fraud Detection, Stolen Check

ACM Reference Format:
Fei Zhao, Jiawen Chen, Bin Huang, Chengcui Zhang, and Gary Warner. 2024. CheckGuard: Advancing Stolen Check Detection with a Cross-Modal Image-Text Benchmark Dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679155

1 Introduction

Financial fraud involving stolen checks poses significant challenges, particularly with platforms such as Telegram being used to sell these checks illicitly. This trend makes it easier for criminals to distribute stolen checks, increasing the need for effective detection and prevention methods. To effectively alert financial institutions or their clients, it is crucial to extract check information from images found on these platforms. Traditional methods, including manual inspections and Optical Character Recognition (OCR) systems, are inadequate. Manual methods are labor-intensive and cannot scale with the rising volume of fraud. Traditional OCR systems struggle with handwritten content variability and depend on predefined check layouts, limiting their effectiveness in accurately detecting and categorizing check information.

The advent of deep learning and multimodal large language models (MLLMs) has shown great promise in various applications, including visual question answering (VQA) [13] and image captioning (IC) [4]. However, these models require extensive training data to achieve high performance on the check information extraction task, which presents a significant hurdle. Current publicly available datasets for check analysis, e.g., [2], are limited in size and scope, restricting the potential of these advanced models.

*Both authors contributed equally to this research.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679155

To address these limitations, we introduce "CheckGuard," a large-scale cross-modal image-to-text dataset designed to train deep learning models, such as MLLMs, for tasks including VQA and IC. Our contributions can be summarized as follows:

1. Novel Cross-Modal Dataset: We present CheckGuard, a comprehensive dataset consisting of 7,245 check image segments annotated with information on **seven key elements**: Drawer, Payee, Date, Amount, Drawee, Routing Number, and Check Number, resulting in **50,715 image-to-text samples**. This dataset offers valuable training data for the community, enabling the development and evaluation of advanced MLLMs for check information extraction.

2. Realistic and Diverse Data: All images are collected from social media platforms used for selling stolen checks. Unlike clean and legit checks obtained using highly standardized scanning templates and protocols, our dataset largely contains check images with different lighting environments, varying degrees of mutilation, and random graffiti. This presents unique challenges but also provides distinct training examples to improve the models' generalization ability. Furthermore, this dataset comprises checks from **more than 15 financial institutions** with diverse layouts and formats, enhancing the robustness of trained models.

3. Benchmarking State-of-the-Art Methods: We evaluate state-of-the-art OCR-based and MLLM-based methods on the CheckGuard dataset. Additionally, we explore potential enhancements to these models by employing parameter-efficient fine-tuning (PEFT) techniques. Our benchmarks provide detailed insights into the performance of these methods across different key elements, highlighting their strengths and limitations.

4. Image Rectification Method: We propose a novel coarse-to-fine image rectification method to correct the orientation of check images during data cleaning and collection. Tilted images hinder the image annotation process and can significantly impact model performance. Our method ensures high-quality data, enhancing the effectiveness of subsequent data annotation and model training.

In summary, CheckGuard addresses the critical need for a large-scale, high-quality dataset in the domain of check information extraction. By providing this resource and benchmarking advanced methods, we aim to drive further advancements in the field, improving the detection and prevention of financial fraud.

2 Related Work

Currently there is no large scale public check image dataset available, let alone one for stolen check images. Existing datasets, such as the IDRBT Cheque Image Dataset [2] and the Bank Check Segmentation Dataset (BCSD) [8], lack the data diversity needed for effective model generalization. The IDRBT Cheque Image Dataset focuses on verifying pen inks of handwritten signatures and contains only 112 checks from four banks in India, scanned at 300 dpi resolution. Similarly, BCSD provides 158 check images with manually labeled segmentation masks for signatures, sourced from the IDRBT dataset and additional anonymized checks from the Internet. These datasets feature idealized data without shadows or distortions, leading to limited effectiveness in training models for real-world applications.

Large multimodal foundation models such as Internvl [1] and BLIP-2 [10] have significantly advanced vision-language tasks but fall short in OCR tasks due to insufficient OCR-specific training data. LLaVA-NeXT [11], designed to excel in handwritten OCR tasks, leverages a pretraining dataset enriched with OCR samples to enhance accuracy. However, the lack of check samples in these pretraining datasets limits their performance in check information extraction. To address this gap, we propose CheckGuard, a comprehensive cross-modal dataset for check information extraction. With a total of 50,715 labeled samples, CheckGuard supports parameter-efficient fine-tuning and the evaluation of advanced MLLMs.

3 CheckGuard Dataset Construction

In this section, we introduce our methods for constructing the CheckGuard dataset, a large cross-modal image-to-text dataset focused on paper check images collected from online platforms that present significant challenges due to their complex layouts, varied formats, and handwritten content. We annotated seven key elements for each check image segment, shown in Fig. 1.

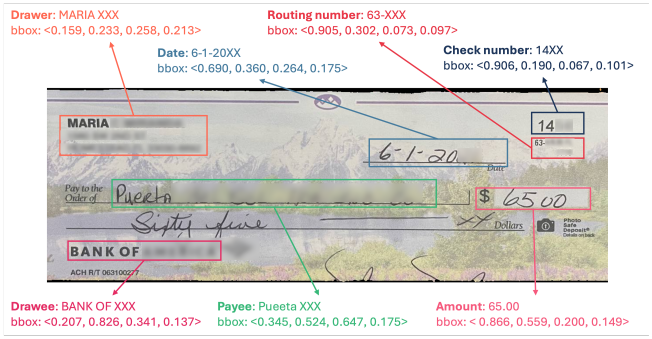

Figure 1: Visualization of the CheckGuard dataset annotations. The image shows a check image segment with annotated elements including Drawer, Payee, Date, Amount, Drawee, Routing Number, and Check Number. Each annotation includes the textual information and the bounding-box coordinates in the format <center x, center y, width, height> (Sensitive details have been blurred to ensure privacy. For full access to the dataset, please contact the authors.)

3.1 Element Definition

Seven key elements are identified from checks, including Drawer, Payee, Date, Amount, Drawee, Routing Number, and Check Number, which play critical roles in financial transactions and fraud detection. For instance:

- **Drawer**: The entity who writes the check, e.g., "John Doe."
- **Payee**: The recipient of the check, e.g., "ABC Company."
- **Date**: The date the check is written, e.g., "05/21/2024."
- **Amount (numerical)**: The monetary value of the check in numerical form, e.g., "$1,234.56."
- **Drawee**: The bank or financial institution where the check can be cashed, e.g., "Bank of America."
- **Routing Number**: The bank's unique identifier, e.g., "12-345/678."
- **Check Number**: A number identifying the check, e.g., "1001."

Additionally, we labeled data for other elements such as the amount in words, memo, signature, and payee address. However, the

amount in words is semantically similar to the numerical amount, and some elements, such as memo and payee address, are less frequently present. Therefore, in this dataset, we only focus on the primary seven elements on checks.

3.2 Dataset Acquisition

All the samples in our dataset are extracted from the real-world images provided by the cyber intelligence company DarkTower. As shown in Fig. 2, the imagery includes scenes of considerable complexity and variability, with checks often overlapping and appearing at different angles. This variability presents significant challenges in automated processing, necessitating check segmentation and rectification for accurate data extraction.

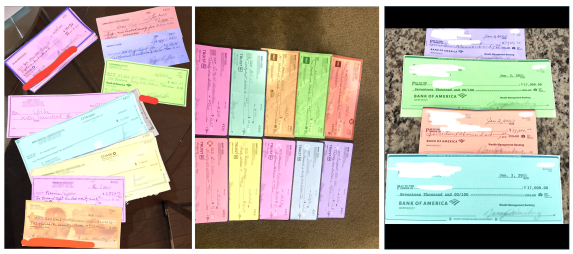

Figure 2: Examples of real-world check images, including overlapping checks and various orientations. Colored areas represent SAM's check segmentation output (Sensitive details have been blurred to ensure privacy. For full access to the dataset, please contact the authors.)

3.3 Check Segmentation and Rectification

To address the challenges presented by the variability in check images, we employed a robust data acquisition and cleaning process. First, all obtained images were processed using the Segment Anything Model (SAM) [9] to isolate potential check image segments. These segments were then classified using a pre-trained YOLO model [7] to identify relevant check segments, ensuring that only check segments proceed to the next stage.

Identified check segments underwent a two-step coarse-to-fine image rectification process to correct skewness and perspective distortions. First, a customized YOLO model [7] is trained to detect 72 rotation degrees, each covering a 5-degree interval, for coarse-level correction. This initial step provides a rough alignment of the check segments. Following this, the Line Segment Detector (LSD) [14] refines the alignment by identifying prominent lines within the image and determining the predominant inclination angle. This comprehensive coarse-to-fine rectification process effectively corrects excessive rotations and aligns the checks accurately. After rectification, the aligned check segments are sent to annotators for labeling. This approach ensures that the CheckGuard dataset is both extensive and of high quality, suitable for training advanced models in check information extraction.

3.4 Data Annotation and Label Format

After segmentation and rectification, we obtained 7,245 check image segments, divided into 6,236 for training, 504 for validation, and 505 for testing. Using the VGG Image Annotator (VIA) tool [3], we meticulously annotated the seven key elements of all segments. Each annotation includes both the textual information and the bounding-box coordinates, formatted for the YOLO model [7], providing the center point (x, y), width, and height as four float values. These annotations are stored in JSON files, where each element type is a key, and the associated ground truth textual labels and bounding-box coordinates are the values. An example is shown in Fig. 1.

In summary, the CheckGuard dataset construction involves element definition, diverse data acquisition, precise image segmentation and rectification, and detailed annotation to create a robust resource for developing advanced check information extraction models.

4 Experiment for Benchmarking

In this section, we benchmark the performance of various models in extracting key textual information from checks. We evaluate the SOTA multimodal large language model (MLLM), LLaVA-Next with Mistral [6] backbone, a contemporary OCR-based method (PaddleOCR [12]), and enhanced versions of these models using YOLO model [7] and parameter-efficient fine-tuning (PEFT) [5].

4.1 Experiment Setup

We conducted experiments using three methods: YOLO-OCR, YOLO-MLLM, and MLLM. Each MLLM model has two versions: zero-shot and PEFT. In the YOLO-based approachs, a YOLO model is trained to detect the 7 elements of interest, and the other part, e.g., OCR or MLLM, processes only these sub-areas of the check image. Therefore, YOLO-based methods only focus on the detected areas of interest, while methods without YOLO process the entire check image directly.

We select Low-Rank Adaptation (LoRA) as our PEFT method. The hyperparameters for LoRA, i.e., LoRA's rank and scaling factor, are set to 128 and 256, respectively. The LoRA layers are applied only on the large language model part of MLLM: Mistral. All the linear layers inside the Mistral will be appended with additional LoRA layers. Each key element extraction task has its own set of LoRA weights trained to optimize performance.

We evaluated the performance of these models across seven key elements: Drawer, Payee, Date, Amount, Drawee, Routing Number, and Check Number. Accuracy (ACC) is measured based on an exact match, requiring a 100% correct extraction. We also use normalized edit distance (NED) [15] to evaluate the performance for certain elements, where a lower NED indicates better performance.

4.2 Results and Analysis

This section presents the results for the seven key elements in the dataset and discusses the challenges and potential research gaps.

4.2.1 Date Element Extraction. As shown in Fig. 3 (a-c), our PEFT-based MLLM achieved over 90% accuracy across all three classes (year, month, and day) for the Date element, with an impressive 97.62% accuracy for the year element. These results demonstrate significant improvements over the OCR-based method and the other MLLM-based models, highlighting the effectiveness of PEFT in enhancing model performance. However, the performance on

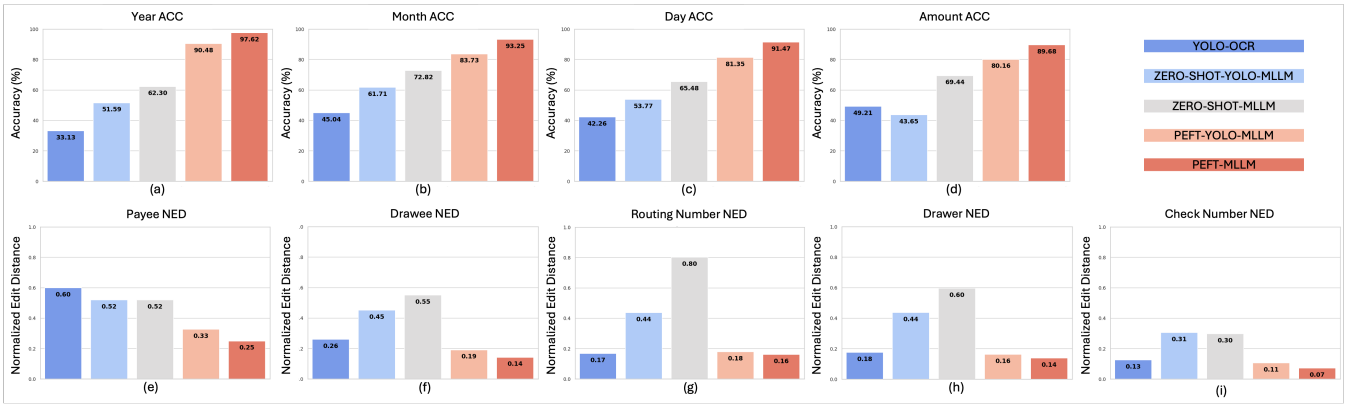

Figure 3: Benchmark results of the experiment

the day component is consistently lower than that on the month and year. This is due to the handwritten "/" being similar to "1", which can confuse the models. For example, the handwritten date "2/2/2024" could be misrecognized as "12" or "21".

4.2.2 Amount Element Extraction. As shown in Fig. 3 (d), the PEFT-based MLLM demonstrates a substantial lead on this task, with an accuracy of 89.68%. This represents significant improvements over both the OCR-based method, enhancing accuracy by 82.24%. A potential research direction here is that, since a check typically has two fields with amount information, i.e., numerical field and textual $ amount field, leveraging both areas could potentially enhance model performance.

4.2.3 Payee Element Extraction. For extracting payee information, the PEFT-based MLLM achieved a marked improvement in normalized edit distance (NED), achieving a value of 0.2494 shown in Fig. 3 (e). This performance substantially surpasses the OCR-based method and the other MLLM-based models, with reductions in NED by 0.3500 and 0.2705 (compared to zero-shot-YOLO-MLLM), respectively. Most of the payee samples are handwritten, which is challenging for OCR-based methods. Since most payees are entities that can be found online, integrating Retrieval-Augmented Generation (RAG) [16] could potentially further improve performance.

4.2.4 Drawee, Routing Number, Check Number, and Drawer Element Extraction. As shown in Fig. 3 (f-i), the PEFT-based MLLM consistently outperforms the other methods, highlighting its robustness. Without PEFT fine-tuning, MLLMs struggle to locate elements of interest due to the similar appearance and formatting of certain elements (e.g., check number and routing number both appearing as integers), which can confuse the models. YOLO-based methods perform better in these cases because YOLO assists in accurately identifying target areas. However, there are risks associated with using this method as incorrect element area detections from the customized YOLO model could introduce noise and potentially negatively impact the performance. Interestingly, YOLO-OCR outperforms zero-shot YOLO-MLLM for Drawee, Routing Number, Check Number, and Drawer elements, but performs worse on Date, Payee, and Amount elements, which are mostly handwritten. This indicates that OCR is better suited for handling printed content, while MLLM excels at processing handwritten content. The results demonstrate the effectiveness of fine-tuning (PEFT) in significantly improving model accuracy for these challenging tasks.

Overall, the PEFT enhancements enable the MLLM to significantly outperform the SOTA methods, setting a new benchmark in the field of automated check analysis.

5 Conclusion and Future Work

This paper presents CheckGuard, a large cross-modal image-to-text dataset designed for the emerging field of paper check fraud detection. The dataset's extensive real-world labeled check data provides a valuable resource for training deep learning models, including multimodal large language models (MLLMs). We established benchmarks on seven key elements and explored potential improvement directions, such as model combination and parameter-efficient fine-tuning (PEFT) on state-of-the-art MLLMs.

In the future, we aim to expand and enhance CheckGuard in several ways. First, we will collect and annotate more real-world data to enrich the dataset. Second, we will explore more sophisticated image rectification methods to improve automation. Third, we plan to investigate MLLM capabilities in key element area detection, enabling the model to generate both content and corresponding locations (e.g., bounding box coordinates) of key elements. Fourth, we intend to reduce the hallucinations of MLLMs by adopting Retrieval-Augmented Generation (RAG) techniques, thereby improving the accuracy and reliability of information extraction. Finally, we intend to streamline the process for dataset access, reducing paperwork and making it more convenient for researchers.

Acknowledgments

The authors would like to thank Ashlynn Schultz from DarkTower for sharing this problem and the raw check image dataset with our team.

This work was supported in part by NSF CNS-2154589, 2154443, and 2154507, "Collaborative Research: SaTC: CORE: Medium: Bubble Aid: Assistive AI to Improve the Robustness and Security of Reading Hand-Marked Ballots," $1,200,000, 10/01/2022-09/30/2026.

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238* (2023).

[2] Prabhat Dansena, Soumen Bag, and Rajarshi Pal. 2017. Differentiating Pen Inks in Handwritten Bank Cheques Using Multi-layer Perceptron. In *Pattern Recognition and Machine Intelligence*, B. Uma Shankar, Kuntal Ghosh, Deba Prasad Mandal, Shubhra Sankar Ray, David Zhang, and Sankar K. Pal (Eds.). Springer International Publishing, Cham, 655–663.

[3] Abhishek Dutta, Ankush Gupta, and Andrew Zissermann. 2016. VGG image annotator (VIA).

[4] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *Comput. Surveys* 56, 3 (2023), 1–39.

[5] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024).

[6] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[7] Glenn Jocher, Abhishek Chaurasia, and Jialian Qiu. 2023. Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics.

[8] Muhammad Saif Ullah Khan. 2021. A novel segmentation dataset for signatures on bank checks. *CoRR* abs/2104.12203 (2021). arXiv:2104.12203 https://arxiv.org/abs/2104.12203

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[12] PaddlePaddle. 2024. PaddleOCR GitHub Repository. https://github.com/PaddlePaddle/PaddleOCR/tree/main. Accessed: May 9, 2024.

[13] Himanshu Sharma and Anand Singh Jalal. 2021. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing* 116 (2021), 104327.

[14] Rafael Grompone Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. 2008. LSD: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence* 32, 4 (2008), 722–732.

[15] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.

[16] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868* (2023).