Innovative Image Fraud Detection with Cross-Sample Anomaly Analysis: The Power of LLMs

Qiwen Wang[1], Junqi Yang[3], Zhenghao Lin[2], Zhenzhe Ying[3], Weiqiang Wang[3], Chen Lin[2]*
[1]Institute of Artificial Intelligence, Xiamen University, China
[2]School of Informatics, Xiamen University, China
[3]Ant Group, China
chenlin@xmu.edu.cn

Abstract

The financial industry faces a substantial workload in verifying document images. Existing methods based on visual features struggle to identify fraudulent document images due to the lack of visual clues on the tampering region. This paper proposes CSIAD (Cross-Sample Image Anomaly Detection) by leveraging LLMs to identify logical inconsistencies in similar images. This novel framework accurately detects forged images with slight tampering traces and explains anomaly detection results. Furthermore, we introduce CrossCred, a new benchmark of real-world fraudulent images with fine-grained manual annotations. Experiments demonstrate that CSIAD outperforms state-of-the-art image fraud detection methods by 79.6% ($F1$) on CrossCred and deployed industrial solutions by 21.7% ($F1$) on business data. The benchmark is available at https://github.com/XMUDM/CSIAD.

1 Introduction

In recent years, almost all users have submitted image versions or screenshots of common documents to financial institutions to authenticate identity, ownership, transaction records, etc. With the advancement of image editing techniques (Chen et al., 2023; Tuo et al., 2024; Zhao et al., 2021; Zeng et al., 2024), malicious parties can generate highly realistic counterfeit materials at minimal cost, posing a serious threat in the financial sector. Automatic *image fraud detection* is critical to enhance security and operational efficiency.

Existing image fraud detection methods predominantly rely on visual features (Zhu et al., 2024; Yu et al., 2024; Guo et al., 2023). However, unlike natural images with complex textures and colors, images encountered in the financial sector are *document images* (Luo et al., 2023; Cloud, 2024), which share a simple background (e.g., typically white), highly structured layouts, and consistent graphical elements (e.g., uniform text fonts). The tampered regions hardly display any visual clues on the edge or surface. As illustrated in Figure 1(a), the SOTA image fraud detection method (Qu et al., 2023) cannot accurately locate the tampered region. Furthermore, it lacks explanations for its decision to speed up human verification.

Although image fraud detection is limited when based on *single-image* visual features, reviewers can detect anomalous images based on *cross-sample* logical inconsistencies in manual verification processes. As illustrated in Figure 1(b), the tampered region includes the account and address information because the forged payment record copies the account number from other images and alters the account name and address.

Since manually cross-sample reasoning is time-consuming, we propose CSIAD, which leverages LLMs for accurate and interpretable image fraud detection. CSIAD operates through two stages: (I) *Retrieval Ensemble*, which discovers potential image associations by aggregating visual and textual retrieval results, and (II) *Fact-Driven Cross-Sample Inference*, which detects and concludes anomaly from associated images. Specifically, in stage II, CSIAD first adopts an LLM to generate *logical rules* that an authentic image should follow, then refines the rules using unsupervised factual evidence, effectively mitigating LLM's hallucinations while improving detection accuracy. Next, CSIAD detects, verifies, and explains the anomalies based on the self-generated rules.

To better address the gaps in image fraud detection in the financial sector, we present a new benchmark CrossCred, different from existing datasets (Dong et al., 2024; Qu et al., 2023; Wang et al., 2022; Qu et al., 2024) in the following aspects. (1) CrossCred is entirely from *real-world* data in complex financial fraud scenarios, where images are meticulously forged to deceive current

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14058–14078

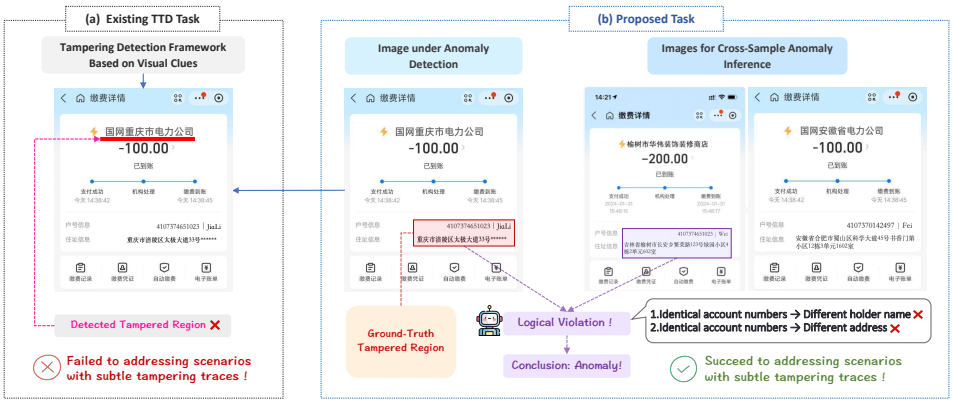

Figure 1: Comparison between TTD (Tampered Text Detection) and CSIAD (Cross-Sample Image Anomaly Detection). The forged image is a mobile payment record. **Left**: Existing TTD relies on visual features of a single sample, which struggles with subtle tampering traces. **Right**: CSIAD incorporates cross-sample analysis based on textual information, effectively increasing the accuracy and explainability of image fraud detection.

detection technology. On the contrary, most existing datasets are synthesized with random modifications. (2) CrossCred contains detailed human annotations, including tampered regions and explanations. (3) CrossCred supports image fraud detection via cross-sample inference (i.e., each fraud case is associated with a few similar images), while previous datasets only contain a single sample for each case.

Experiments on CrossCred show that CSIAD improves the image fraud detection performance, achieving an $F1$ of 81.9%, outperforming the SOTA image tampering detection method by 79.6%. Further evaluation on a real-world business dataset reveals that our model achieves an $F1$ of 85.7%, surpassing deployed industrial solutions by 21.7%.

In conclusion, our study makes several noteworthy contributions:

- **Novel methodology**. We present CSIAD: the first model that detects counterfeit document images by leveraging LLMs to uncover logical inconsistencies through cross-sample inference. In addition to enhancing detection accuracy, CSIAD provides fine-grained explanations that enable real-time monitoring and analysis of anomalous traffic.
- **New benchmark**. We release CrossCred: a real-world dataset that covers various types of fraud images in the financial sector with detailed human annotations to evaluate image fraud detection methods.
- **Superior performance in academic and business scenarios**. CSIAD achieves accurate and interpretable fine-grained anomaly detection on both CrossCred and industrial datasets, showing significant improvements over academic baselines and sophisticated business solutions.

2 Related Work

Document Image Tampering Datasets. Recent datasets like RIFLC (Cloud, 2022), TTI (Cloud, 2023), TextTamper (Dong et al., 2024), DocTamper (Qu et al., 2023), Tampered-IC13 (Wang et al., 2022), and OSTF (Qu et al., 2024) have been developed for tampering detection. However, they focus on single-sample analysis, lack fine-grained manual annotations, and fail to replicate real-world fraud scenarios. In contrast, our proposed CrossCred features fine-grained manual annotations and supports cross-sample detection.

Document Image Fraud Detection Methods. Existing methods for document image fraud detection focus on visual element anomalies at the individual sample level (Qu et al., 2023; Sun et al., 2024; Wang et al., 2022; Dong et al., 2024; Xu et al., 2022; Qu et al., 2024; Sun et al., 2019). However, they rely on obvious visual cues and struggle with subtle tampering traces, revealing their limitations. In this study, we propose CSIAD, leveraging LLMs for cross-sample analysis to complement existing approaches.

Self-Reflection Ability of LLMs. Recent research on self-reflection introduces advanced prompting strategies, enabling LLMs to refine responses via feedback (Shinn et al., 2023; Bai et al.,

2022; Paul et al., 2024; Madaan et al., 2023). Feedback can come from external sources (e.g., models, tools, knowledge bases) (Gou et al., 2024; Olausson et al., 2024; Chen et al., 2024; Nathani et al., 2023; Kim et al., 2023) or internal self-assessment (Madaan et al., 2023; Zhang et al., 2024; Yao et al., 2024; Lightman et al., 2024). However, studies (Huang et al., 2024; Zhang et al., 2024; Stechly et al., 2023; Liang et al.) show LLMs' correction capabilities are unreliable without external feedback. In this study, we propose a fact-driven verification mechanism, guiding LLMs to autonomously select reflection paths based on objective factual data, ensuring reliable outputs.

3 Methodology

3.1 Framework Overview

To determine whether a document image is fraudulent, a natural workflow is to "consult" relevant images to identify the potential anomalies. As shown in Figure 2, `CSIAD` contains four modules: the *Retrieval Ensemble* module aims to recall as many relevant images given the query image I_q to form a batch of suspicious images $\mathcal{S}_q$; the *Sample-Specific Rule Generation* module generates a set of rules $\mathcal{R}_{ini}$ that normal images must adhere to; the *Fact-Driven Verification* module further validates and filters the rules to obtain $\mathcal{R}_{ver}$; and finally, the *Rule-based Anomaly Detection* module detects and interprets the anomalies in $\mathcal{S}_q$.

3.2 Retrieval Ensemble

We combine visual and textual retrieval to maximize the recall of relevant images. Adopting visual retrieval alone can lead to increased computational complexity by analyzing irrelevant images, e.g., certificates may share similar visual appearances yet contain different information (Figure 3(a)). It can also reduce detection accuracy by missing relevant images, e.g., images differ visually while conveying related content (Figure 3(b)).

Furthermore, we combine dense retrieval and sparse retrieval. The former captures document-level semantic relationships, while the latter excels at nuanced keyword connections, benefiting highly structured documents.

Specifically, the visual features are extracted by CN-CLIP (ViT-B/16) (Yang et al., 2023), dense textual vectors by CN-CLIP (RoBERTa-wwm-Base) (Yang et al., 2023). We use an LLM to extract structured texts $\mathcal{T}$ and a HashingVectorizer to extract sparse textual vectors.

The similarity score between the given image I_q and any image in the database I_i is computed by:

$$\begin{aligned} s(I_q, I_i) = \alpha \times \cos(\mathbf{v}_q, \mathbf{v}_i) \\ + \beta \times \cos(\mathbf{e}_q, \mathbf{e}_i) \\ + \gamma \times \cos(\mathbf{t}_q, \mathbf{t}_i), \end{aligned} \tag{1}$$

where $\mathbf{v}_q, \mathbf{e}_q, \mathbf{t}_q$ are the visual feature, dense textual vector, and sparse textual vector, respectively. $\cos(\cdot, \cdot)$ denotes the cosine similarity. The weighting coefficients α, β, and γ are meticulously calibrated through preliminary experiments.

Finally, the retrieved images associated with I_q are collected to form $\mathcal{S}_q$:

$$\mathcal{S}_q = \{I_i \mid \forall s(I_q, I_i) \geq \delta\}, \tag{2}$$

where δ is the correlation truncation score.

3.3 Sample-Specific Rule Generation

The retrieved batch of images $\mathcal{S}_q$ shows high similarity, suggesting potential anomalies. Intuitively, we can use an LLM $\mathcal{M}$ to infer abnormal images based on a set of rules. The rules can be predefined, e.g., by human experts. Unfortunately, since new types of documents emerge, e.g., payment via a newly launched app, it is impractical to enumerate all possible rules manually. An LLM can explicitly or implicitly generate the rules. However, if no groundings are provided, the LLM often generates trivial rules, e.g., *the payment amount should be negative*.

Consequently, we encourage $\mathcal{M}$ to generate sample-specific rules. As shown in Figure 2(b), we provide $\mathcal{M}$ with structured text $\mathcal{T}$ (details in the last subsection) and few-shot business rule examples $\mathcal{F}_{\text{rule}}$. Given the potentially lengthy context of $\mathcal{T}$, the first step in the prompt $p_{\text{gen\_rule}}$ for Sample-Specific Rule Generation is to focus on key fields within $\mathcal{T}$. Leveraging these key fields and $\mathcal{F}_{\text{rule}}$, we then facilitate the generation of initial rules $\mathcal{R}_{ini}$.

$$\mathcal{R}_{ini} = \mathcal{M}(p_{\text{gen\_rule}}(\mathcal{T}, \mathcal{F}_{\text{rule}})) \tag{3}$$

To ensure that edge anomaly cases such as unconventional or emerging anomaly patterns are also detected, we incorporate a temperature sampling strategy (Chang et al., 2023), which enables $\mathcal{M}$ to generate more creative rules during decoding by utilizing a higher sampling temperature.

3.4 Fact-Driven Rule Verification

We implement a fact-based verification mechanism to tackle potential errors or hallucinations in $\mathcal{R}_{ini}$,

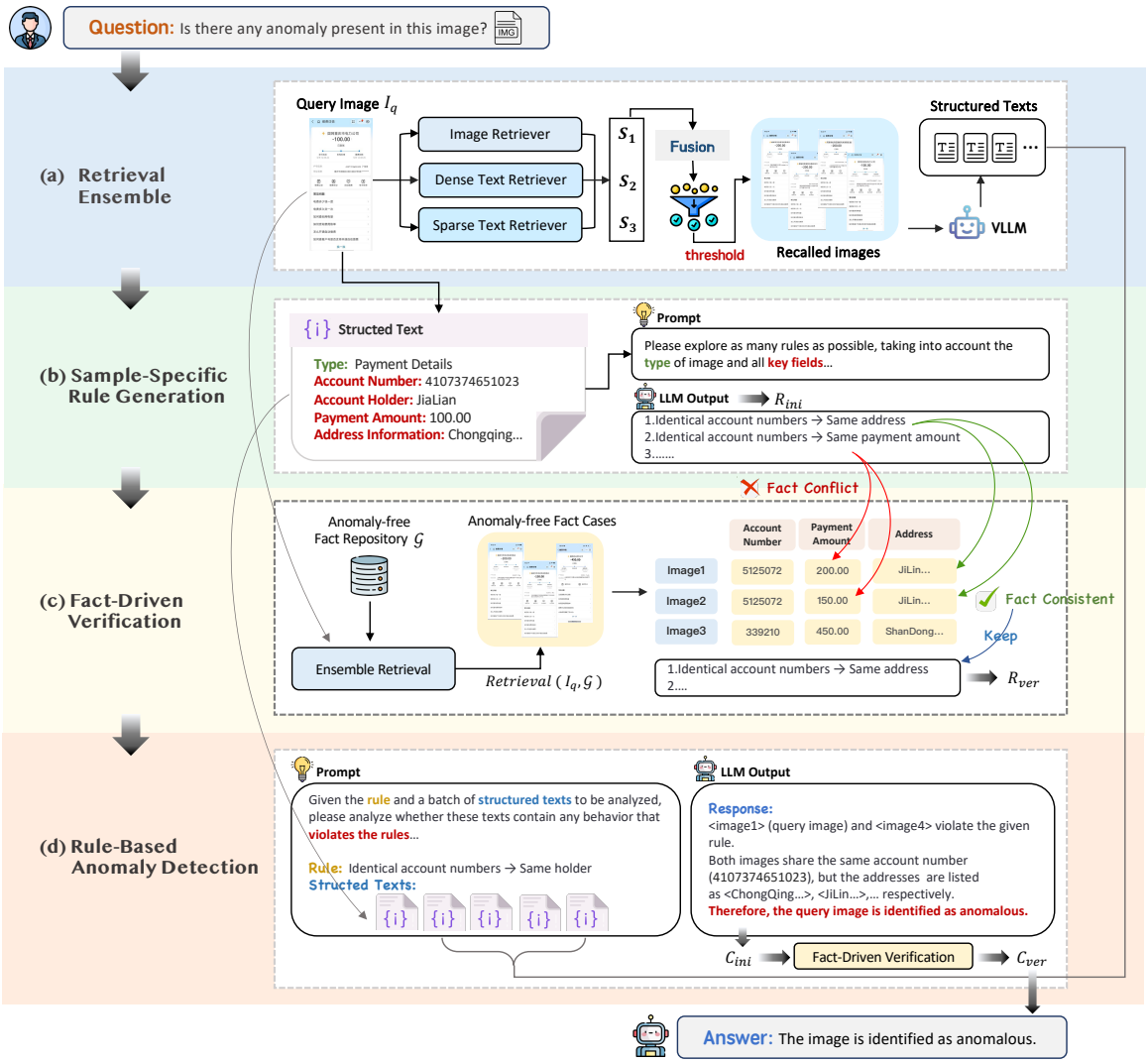

Figure 2: Framework for CSIAD. The diagram outlines the key modules: Retrieval Ensemble, Sample-Specific Rule Generation, Fact-Driven Verification, and Rule-Based Anomaly Analysis.

e.g., *"Identical taxpayer identification number → Same tax paid"*. The assumption is that if we hold a fact repository that covers a sufficient number of scenarios, we can detect and remove the erroneous or hallucinatory rules.

Specifically, we construct an anomaly-free fact repository $\mathcal{G}$. Our key insight is that the quality of rule verification improves as the size of $\mathcal{G}$ increases, as a larger repository covers a broader range of scenarios described by the rules, thus enhancing the comprehensiveness of the verification. $\mathcal{G}$ consists of historical images submitted in the system. Firstly, real-world business databases are typically vast, with an inherently low anomaly rate (approximately one in a thousand). Secondly, the anomaly rate can be further reduced by employing basic business anomaly detection methods to filter out obvious outliers, making $\mathcal{G}$ reliable for ground truth fact verification.

We then use $\mathcal{G}$ to validate the rules in $\mathcal{R}_{ini}$. Formally, we first retrieve factual images from $\mathcal{G}$ that match I_q:

$$\mathcal{G}_q = \{I_g^{(i)} \mid \forall I_g^{(i)} \in \mathcal{G},\ s(I_q, I_g^{(i)}) \geq \delta\} \tag{4}$$

Then, for $r_{\text{ini}}^{(i)} \in \mathcal{R}_{ini}$ to be validated, we employ a prompt $p_{\text{ver\_rule}}$ designed for fact-driven rule verification, guiding $\mathcal{M}$ to examine $\mathcal{G}_q$:

$$\begin{aligned}\mathcal{R}_{ver} = \{r_{\text{ini}}^{(i)} \mid \forall r_{\text{ini}}^{(i)} \in \mathcal{R}_{ini},\\ \mathcal{M}(p_{\text{ver\_rule}}(r_{\text{ini}}^{(i)}, \mathcal{G}_q)) = \text{True}\}\end{aligned} \tag{5}$$

If $\mathcal{M}$ searches for any instances where the conditions specified by $r_{\text{ini}}^{(i)}$ are violated in $\mathcal{G}_q$, it signifies that $r_{\text{ini}}^{(i)}$ is not universally applicable to all valid images and is subsequently pruned.

Figure 3: Visual and Textual Discrepancies in Financial Document Images. (a) Two visually similar financial document images with entirely different textual content. (b) Two visually dissimilar financial document images with highly related textual content.

3.5 Anomaly Detection and Verification

Once a set of fact-grounded rules is established, the next step is to leverage $\mathcal{R}_{ver}$ to guide the anomaly detection phase(Figure 2(d)).

Rule-Based Anomaly Detection. We let $\mathcal{M}$ to sequentially verify each rule in $\mathcal{R}_{ver}$, i.e., whether the given image group $\mathcal{S}_q$ exhibits rule violations through prompt p_{gen_con}. We denote the initial anomaly conclusions as:

$$\mathcal{C}_{ini} = \{\mathcal{M}(p_{gen\_con}(\mathcal{T}, r_{ver}^{(i)}))| \, \forall r_{ver}^{(i)} \in \mathcal{R}_{ver}\} \quad (6)$$

each anomaly conclusion $c_{ini}^{(i)} \in \mathcal{C}_{ini}$ includes the relevant image IDs and the explanation that details the cause of the anomaly and its location.

Fact-Driven Anomaly Verification. Given the lengthy and complex contexts within the structured text of multiple images, there is an elevated risk of factual errors in $\mathcal{C}_{ini}$. To ensure reliable anomaly detection, it is essential to verify $\mathcal{C}_{ini}$. During this verification phase, the fact repository denotes the structured textual $\mathcal{T}$ furnished to $\mathcal{M}$. We assess the validity of each conclusion to determine whether it is a degenerated case due to insufficient supporting evidence in $\mathcal{T}$ through a specifically designed prompt p_{ver_con}. The verified anomaly conclusions are denoted as $\mathcal{C}_{ver}$.

$$\begin{aligned}\mathcal{C}_{ver} = &\{c_{ini}^{(i)} \mid \forall c_{ini}^{(i)} \in \mathcal{C}_{ini}, \\ &\mathcal{M}(p_{ver\_con}(\mathcal{T}, c_{ini}^{(i)})) = \text{True}\}.\end{aligned} \quad (7)$$

Table 1: Comparison between CrossCred and existing image tampering detection datasets (DocTamper (Qu et al., 2023), TextTamper (Dong et al., 2024)).

Aspect	DocTamper	TextTamper	CrossCred
Data Source	Simulation	Simulation	Real-world
Tamper	Random	Random	Targeted
# Doc Type	N/A	N/A	61
Annotation	Coarse	Coarse	Fine-grained
Detection	Single-Sample	Single-Sample	Multi-Sample

For the concrete prompts utilized in each module, please refer to Appendix D.

4 Benchmark

As shown in Table 1, existing document image anomaly detection datasets synthesize images by tampering with a random region, e.g., copy-move, splicing, etc. They fail to simulate real-world targeted image fraud, such as altering critical fields like amounts, dates, and identity information. They lack fine-grained manual annotations. Additionally, they only contain one sample for an abnormal image, forcing the detection method to rely purely on the visual features of a single sample. While forged images are produced in batches nowadays, single-sample detection is suboptimal.

To address these shortcomings, we propose a benchmark consisting of real-world fraud images, where the tampering is proficiently targeted for financial theft and crimes. The benchmark[1] contains fine-grained, manually annotated anomalies, and supports multi-sample detection.

4.1 Benchmark Construction

The benchmark construction process consists of four steps, as illustrated in Figure 4.

Step 1: Image Grouping. In the financial service industry, diverse document types are utilized, such as tax invoices, financial statements, identity documents, contracts, etc. Each document type can be associated with various formats, e.g., mobile payment records by different apps. Images of the same document type and format are grouped together to form a candidate *case* for analysis.

[1]Data strictly desensitized and encrypted with comprehensive protection measures, containing no Personally Identifiable Information.

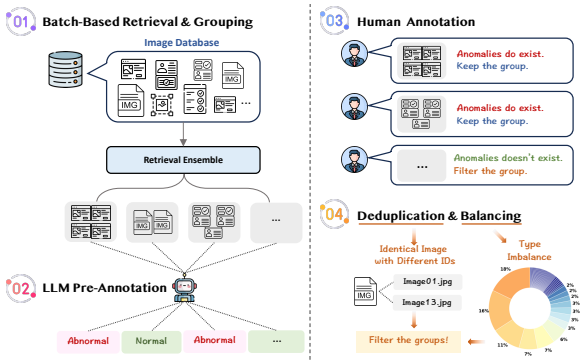

Figure 4: The construction process of CrossCred.

Step 2: Pre-Annotation. It is infeasible to check all images manually, given the large volume. Thus, we leverage an LLM to pre-annotate the image groups formed in Step 1. If anomalous behaviors are detected among the images within a group, we classify them as an anomalous case and retain them for further analysis. Specifically, from a collection of 400K business images, we identified 4,063 potential anomalous cases.

Step 3: Human Annotation. Our analysis indicates that pre-annotated anomalous cases contain a notable proportion of false positives. To enhance reliability, we manually verify these cases through expert evaluation, involving (1) logical consistency checks against financial rules and (2) commonsense validity assessments. Each case undergoes $\geq$2-minute expert review, resulting in 558 validated anomalous cases. Each case includes multiple images of the same document type and format. Detailed annotations identify the (possibly multiple) anomalous points in each image, specifying their locations and interpretations.

Our data annotation work was carried out by a professional annotation team within the company. This team consists of 10 experienced annotators and follows a strict three-level quality control process: 1) Preliminary annotation by annotators. 2) Double-checking by a quality inspection team leader and the annotation project leader. 3) Final confirmation by the data requester. For further details on the annotation process, see Appendix G.

Step 4: Deduplication and Balancing. We observed frequent resubmissions of images (e.g., through different URLs) and imbalanced distributions of document types. To address this, we performed deduplication to eliminate redundant images and balanced the document types to ensure uniform distribution, resulting in the retention of 109 anomalous cases.

4.2 Benchmark Statistics

Our benchmark comprises 109 cross-sample anomaly cases (396 total samples, average of 3.63 samples per case) and 109 randomly selected anomaly-free images, totaling 505 images, spanning 61 document types. The complete type distribution of CrossCred is detailed in Appendix F. We plan to release the CrossCred benchmark upon acceptance.

5 Experiments

5.1 Implementation details

Settings. For the experiments, we use two LLMs: Qwen2-VL (Wang et al., 2024) for structured text extraction and Qwen2.5-72B-Instruct (Qwen et al., 2025) for the rest of the tasks. Except for the 1.0 temperature used when sampling the creative rules, we use a temperature of 0.0 when calling the API. The retrieval module adopts weights $\alpha = 0.5$, $\beta = 0.1$, and $\gamma = 0.4$, which achieve superior relevant sample recall performance (Figure **??**). The correlation truncation score δ is set to 0.88.

Evaluation Set. For evaluation set construction, we adopt a sampling approach: one image per anomaly case is allocated to the evaluation set as I_q, and the remaining images in the case are included in the database for retrieval. This results in a balanced evaluation set containing 218 images (109 normal / 109 anomalous), while the reference database comprises 98,758 images.

Evaluation Metrics. Given that each anomaly case may contain multiple anomalies, we implement dual-level evaluation metrics (coarse/fine-grained) to assess the performance.

Coarse-grained metrics. The primary objective is to determine whether an image is correctly identified as anomalous. We treat anomaly detection as a binary classification task, evaluated through standard measures, including Accuracy (Acc), Recall (Rec), Precision (Pre), and F1 score ($F1$).

Fine-grained metrics. These are designed to verify whether multiple anomalies are correctly located and explained in each case. Specifically, an LLM is called to compare $\mathcal{C}_{ver}$ with ground-truth annotations. Fine-grained accuracy is calculated based on whether the localization and explanation

Table 2: Evaluation results on CrossCred dataset. *Retriever* → *Single* refers to retrieving relevant images using only the image modality, while *Retriever* → *Ensemble* refers to retrieving images using ensemble retrieval. Methods marked with * indicate the use of BizRule for reflection.

Method	Coarse-grained				Fine-grained		
	Acc	*Rec*	*Pre*	*F1*	$Rec^{\dagger}$	$Pre^{\dagger}$	$F1^{\dagger}$
DTD	48.6	43.1	48.5	45.6	-	-	-
Tifdm	49.1	34.9	48.7	40.6	-	-	-
Retriever → Single							
CoT Prompt	29.8	35.8	32.0	33.8	10.2	31.3	15.4
Self-Consistency*	76.6	68.8	81.5	74.6	28.2	45.7	34.9
Self-Check*	74.3	72.5	75.2	73.8	29.9	47.9	36.8
Self-Reflection*	74.8	74.3	75.0	74.7	29.6	47.9	36.6
CSIAD	78.0	75.2	79.6	77.4	30.2	50.5	37.8
Retriever → Ensemble							
CoT Prompt	31.7	36.7	33.3	34.9	9.8	35.8	15.4
Self-Consistency*	77.1	69.7	81.7	75.3	28.6	48.2	35.9
Self-Check*	72.5	73.4	72.1	72.7	29.9	49.5	37.3
Self-Reflection*	74.3	77.1	73.0	75.0	30.6	53.4	38.9
CSIAD	**82.1**	**80.7**	**83.0**	**81.9**	**32.1**	**55.1**	**40.6**

of $c_{ver}^{(i)}$ match the annotations. The following aspects are emphasized: Recall ($Rec^{\dagger}$), Precision ($Pre^{\dagger}$), and F1 score ($F1^{\dagger}$). For a more detailed calculation procedure of the fine-grained metrics, please refer to Appendix B.

5.2 Baseline

We conducted a comparative analysis with state-of-the-art document image tampering detection approaches in the field of computer vision, including DTD (Qu et al., 2023) and Tifdm (Dong et al., 2024).

We also evaluate CSIAD against several baselines by applying advanced LLM inference strategies on CSIAD's retrieved relevant images. These baselines include the Standard CoT Prompt (Kojima et al., 2022), Self-Consistency (Wang et al., 2023), Self-Reflection (Shinn et al., 2023), and Self-Check (Miao et al., 2024). Since methods incorporating reflection processes are not consistently reliable without external knowledge (Huang et al., 2024; Zhang et al., 2024), we offer a manually curated rule library (BizRule) from the production system to serve as a reference for reflection.

5.3 Main Results

Table 2 summarizes the experimental results on CrossCred, demonstrating CSIAD's superiority across all baselines. Due to the lack of publicly available real-world datasets from similar scenarios, this paper does not include a comparison across different open benchmark datasets.

Table 3: Comparison of the recall rates between *Retriever* → *Single* and *Retriever* → *Ensemble* in retrieving relevant images.

Method	*Recall*@1	*Recall*@3	*Recall*@5
Retriever → Single	28.2	62.4	77.7
Retriever → Ensemble	**33.8**	**73.2**	**90.6**

Compared with traditional image tampering detection methods, their performance in scenarios lacking evident tampering traces is suboptimal (Table 2). In contrast, CSIAD significantly outperforms these methods, achieving average $F1$ improvements of 80.2% and 90.1% in single-modal retrieval and ensemble retrieval, respectively. This highlights the effectiveness of CSIAD in addressing challenging cases where traditional methods fail.

Table 2 also shows that using the ensemble retriever significantly improves Rec across all evaluated methods and boosts all metrics for CSIAD. To further explore the impact of retrieval ensemble, we conducted a comparative analysis between the single-modal retriever and the ensemble retriever in retrieving relevant images. We annotate the relevant images and report the recall rates in Table 3. The ensemble retriever notably increases $Recall@1$, $Recall@3$, and $Recall@5$. Thus, fusing visual and textual retrieval results can obtain more relevant images, and improve the recall rate of anomaly detection in Table 2.

At the same time, the improved recall of relevant images leads to decreased Acc and Pre for some baselines because it is more challenging to identify image fraud in a larger set. On the contrary, CSIAD leverages a fact-driven verification mechanism that dynamically allows the LLM to adapt its verification path based on specific queries. This adaptability significantly improves both Acc and Pre, with gains of 5.3% and 4.3%, respectively.

Compared with other baselines using LLM inference strategies, our method achieves optimal fine-grained performance, attaining an $F1^{\dagger}$ of 40.6%, a 4.4% improvement over the best baseline. This gain stems from the fact-driven verification mechanism, which grounds generated rules and anomaly conclusions with objective factual data rather than subjective model inferences, effectively reducing fine-grained misjudgments.

Table 4: The Effect of the Different Modules. We conduct comparisons across four modules: Temperature Sampling(TS), Rule Generation(RG),Rule Verification(RV),and Anomaly Verification(AV)

Methods			Coarse-grained				Fine-grained		
RG	RV	AV	Acc	Rec	Pre	$F1$	$Rec^{\dagger}$	$Pre^{\dagger}$	$F1^{\dagger}$
✘	✘	✘	31.7	36.7	33.3	34.9	9.8	35.8	15.4
w/o TS									
✔	✘	✘	61.9	79.8	58.8	67.7	33.7	44.7	38.4
✔	✔	✘	79.4	77.1	80.8	78.9	31.9	51.0	39.3
✔	✘	✔	61.9	76.2	60.3	66.1	31.9	46.1	37.7
✔	✔	✔	79.8	75.2	82.8	78.9	30.2	53.7	38.7
w/ TS									
✔	✘	✘	60.6	82.6	57.3	67.7	34.6	45.2	39.2
✔	✔	✘	81.2	81.7	80.9	81.3	33.0	52.0	40.4
✔	✘	✔	62.4	**82.6**	58.8	68.7	**34.6**	47.6	40.1
✔	✔	✔	**82.1**	80.7	**83.0**	**81.9**	32.1	**55.1**	**40.6**

5.4 Ablation Analysis

Effectiveness of Temperature Sampling. We introduced a temperature sampling strategy during the rule generation phase, utilizing a high temperature ($T = 1.0$). Table 4 shows that $\mathcal{M}$'s performance with TS outperforms that without TS ($T = 0.0$). In particular, Rec increased by 7.31%, reaching 80.7%. This phenomenon arises because $T = 0.0$ produces deterministic rules, which are insufficient for the probabilistic nature of real-world image diversity. In contrast, higher-temperature sampling ($T > 0.0$) expands $\mathcal{M}$'s probability space, enabling the generation of diversified rules that capture edge anomalies more effectively than fixed-pattern ones.

Effectiveness of Rule Verification Modules. Table 4 also shows that removing the rule verification process—where $\mathcal{M}$ reflects on anomalies based solely on initial rules—improves recall but significantly reduces precision, with Pre dropping by 14.2% (with TS) and 13.6% (without TS). We conclude that well-defined rules provide $\mathcal{M}$ with clear analytical guidance, enabling it to focus more effectively on identifying and interpreting critical information.

Quality of `CSIAD` 's Reflection. We further analyzed the quality of rule reflection results in Table 5. Based on the correctness of pre- and post-verification results, all reflective processes were categorized into three types: (1) **Invalid Reflection** (✘ ⇒ ✘): both pre- and post-reflection results are incorrect; (2) **Toxic Reflection** (✔ ⇒ ✘): an initially correct response becomes incorrect after reflection; (3) **Valid Reflection** (✘ ⇒ ✔): an initially incorrect conclusion is corrected.

Table 5: Comparison of reflection outcomes across different LLM-based reasoning strategies. We categorize each case based on the correctness of the pre- and post-reflection outputs: Invalid, Toxic, and Valid.

Method	Invalid↓ ✘ ⇒ ✘	Toxic↓ ✔ ⇒ ✘	Valid↑ ✘ ⇒ ✔
Self-Consistency	**40.4%**	5.2%	54.4%
Self-Check	42.6%	3.7%	53.7%
Self-Reflection	44.1%	4.00%	51.9%
`CSIAD`	43.4%	**2.1%**	**54.5%**

Table 6: Comparison between the deployed industrial solution and `CSIAD` on both the proposed benchmark (`CrossCred`) and real-world business data (Biz.Data).

Dataset	Method	Coarse-grained			
		Acc	Rec	Pre	$F1$
`CrossCred`	Deployed	59.6	28.4	75.6	41.3
	`CSIAD`	**82.1**	**80.7**	**83.0**	**81.9**
Biz.Data	Deployed	84.3	67.9	73.1	70.4
	`CSIAD`	**92.0**	**85.7**	**85.7**	**85.7**

The statistics for each reflection category across different methods are presented in Table 5. Our analysis demonstrates that `CSIAD` optimizes rule verification in two aspects: boosting Valid Reflection rates (correcting errors while preserving valid rules) and reducing Toxic Reflection rates (erroneously filtering out valid rules). Maintaining accurate rules is crucial in anomaly detection, and `CSIAD` strikes an optimal balance between correction and retention. These results highlight how `CSIAD` utilizes reflection to improve both accuracy (by fixing errors) and reliability (by reducing over-correction).

5.5 Comparative Analysis with Existing Industrial Solution

To compare `CSIAD` with industrial solutions[2], we evaluate the latter on `CrossCred` and validate `CSIAD` using industry-confirmed anomalous cases. Experimental results are presented in Table 6.

Performance of Industrial Solution on `CrossCred`. We conducted a comparative evaluation of `CSIAD` against the solution deployed in business systems. The experimental results, as shown in Table 6, demonstrate that `CSIAD` significantly outperforms the existing industrial solution

[2]Due to security protocols, the methodology behind business solutions remains confidential and is not publicly disclosed.

on CrossCred, achieving a 98.3% improvement in $F1$. This underscores the superior effectiveness of CSIAD in handling the CSIAD task.

Performance of CSIAD on Existing Business Data. CrossCred is specifically designed for the CSIAD task. To provide a more comprehensive assessment of CSIAD's performance in real-world business scenarios, we collected a batch of real-world cases confirmed as anomalous by industrial practices for further evaluation. The experimental results reveal that CSIAD achieves significant performance in these real-world scenarios, with a 21.7% improvement in $F1$, further demonstrating its effectiveness in handling complex, real-world cases.

6 Conclusion

In this work, we introduce CSIAD, the first method to leverage LLM reasoning for cross-sample logical anomaly detection in image fraud detection, showcasing the potential of using textual features to uncover image anomalies. We propose an innovative Fact-Driven Verification mechanism as a post-generation refinement method to address potential errors or hallucinations by LLMs, enabling LLM reflections to autonomously select reasoning paths and focus on objective factual data, effectively reducing misjudgments. Experimental results demonstrate that CSIAD significantly outperforms traditional image tampering detection methods and existing industrial solutions. Given these advantages, we believe our framework can benefit various real-world applications.

7 Acknowledgments

Chen Lin is the corresponding author. Chen Lin is sponsored by the Natural Science Foundation of China (No.62372390,62432011) and CAAI-Ant Group Research Fund.

8 Limitations

Our research primarily focuses on images of documents encountered in financial business scenarios. Despite the promising results demonstrated in this study, our method has several limitations. Firstly, we only have access to images within a limited time frame. Although we have comprehensively mined and showcased the existing anomaly types and image categories in the CrossCred benchmark, our research scope remains constrained by the available data. Additionally, due to privacy and security considerations, the scale of our benchmark is restricted. If future access to a broader image database becomes available, we will be able to acquire a more diverse and research-worthy set of anomaly cases. Secondly, one potential limitation of CSIAD is its reliance on the instruction-following capabilities of LLMs. While CSIAD is designed to leverage advanced LLMs for cross-sample analysis, the use of smaller or less capable LLMs may result in suboptimal performance, as their ability to interpret and execute complex instructions is comparatively weaker. Solutions such as enhancing the instruction-following capabilities of smaller LLMs through fine-tuning or leveraging knowledge distillation techniques to transfer the capabilities of larger models to smaller ones could be explored.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. 2023. Kl-divergence guided temperature sampling. *Preprint*, arXiv:2306.01286.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023. Textdiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems*, volume 36, pages 9353–9387. Curran Associates, Inc.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

Alibaba Cloud. 2022. Real-world image forgery localization challenge.

Alibaba Cloud. 2023. Detecting tampered text in images tianchi competition.

Alibaba Cloud. 2024. Ai identity verification - financial certificate tampering detection.

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

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 39648–39677. Curran Associates, Inc.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Dongliang Luo, Yu Zhou, Rui Yang, Yuliang Liu, Xianjin Liu, Jishen Zeng, Enming Zhang, Biao Yang, Ziming Huang, Lianwen Jin, and Xiang Bai. 2023. Icdar 2023 competition on detecting tampered text in images. In *Document Analysis and Recognition - ICDAR 2023*, pages 587–600, Cham. Springer Nature Switzerland.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations*.

Deepak Nathani, David Wang, Liangming Pan, and William Wang. 2023. MAF: Multi-aspect feedback for improving reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6591–6616.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2024. Is self-repair a silver bullet for code generation? *Preprint*, arXiv:2306.09896.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5937–5946.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *Preprint*, arXiv:2310.12397.

Hao Sun, Jie Cao, Zhida Zhang, Tao Wu, Kai Zhou, and Huaibo Huang. 2024. Learning fine-grained and semantically aware mamba representations for tampered text detection in images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 44–57. Springer.

Keyu Sun, Gang Cao, Qian Zhao, and Jianglong Zhang. 2019. Differential abnormality-based tampering detection in digital document images. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pages 145–149.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2024. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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.

Wenbo Xu, Junwei Luo, Chuntao Zhu, Wei Lu, Jinhua Zeng, Shaopei Shi, and Cong Lin. 2022. Document images forgery localization using a two-stream network. *International Journal of Intelligent Systems*, 37(8):5272–5289.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2023. Chinese clip: Contrastive vision-language pretraining in chinese. *Preprint*, arXiv:2211.01335.

Yuxuan Yao, Han Wu, Zhijiang Guo, Biyan Zhou, Jiahui Gao, Sichun Luo, Hanxu Hou, Xiaojin Fu, and Linqi Song. 2024. Learning from correctness without prompting makes llm efficient reasoner. *Preprint*, arXiv:2403.19094.

Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. 2024. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12765–12774.

Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. 2024. Textctrl: Diffusion-based scene text editing with prior guidance control. *Preprint*, arXiv:2410.10133.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622. Association for Computational Linguistics.

Lin Zhao, Changsheng Chen, and Jiwu Huang. 2021. Deep learning-based forgery attack on document images. *IEEE Transactions on Image Processing*, 30:7964–7979.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Jizhe Zhou. 2024. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. *Preprint*, arXiv:2412.13753.

A Distribution of Multi-Feature Fusion Recall Scores.

We constructed a dedicated validation set to optimize recall parameters. Using a grid search approach, we explored the optimal weight coefficients assigned to three features: image features, dense textual features, and sparse textual features. The comprehensive findings from this experiment are illustrated in Figure **??**.

Observations reveals that the optimal weight combination was determined to be $[0.5, 0.1, 0.4]$, corresponding to weight coefficients of $\alpha = 0.5, \beta = 0.1, \gamma = 0.4$. This result highlights that sparse textual features play a crucial role in improving recall accuracy, with importance comparable to image features. This aligns with our analysis: in scenarios with a high density of similar certificate images, keywords or key terms within certificates are critical for differentiation.

B Regarding Fine-grained Metric Calculation

Necessity for fine-grained metrics. A set of financial documents to be analyzed may contain multiple anomalies (e.g., in the set of ID cards, there could be three independent anomalies: *same ID number but different names / different birth dates / different addresses*). Therefore, we have conducted fine-grained individual annotations for each anomaly point in every anomalous case within the benchmark, rather than simply labeling as *anomalous*.

Calculation of fine-grained metrics. For a given anomaly case, the ground truth annotation is $L = \{l_1, l_2, \ldots, l_k\}$, where l_i represents the i-th anomaly point. Let the CSIAD's predicted anomalies be $P = \{p_1, p_2, \ldots, p_m\}$.

To determine semantic equivalence between predicted and ground-truth anomalies, we define a binary matching function:

- Match(p, l) = 1 if the predicted anomaly p and the ground-truth anomaly l are semantically equivalent;
- Match(p, l) = 0 otherwise.

For each predicted anomaly p_j in the prediction set P, we compare it against all ground-truth anomalies l_i in the annotation set L.

- If there exists at least one ground-truth anomaly l_i such that Match$(p_j, l_i) = 1$, then p_j is considered a valid prediction, meaning that the LLM has successfully identified an anomaly that is semantically equivalent to a labeled ground truth.
- If no such l_i is found, then p_j is counted as a false positive.

To quantitatively evaluate fine-grained semantic correctness, we define the following metrics.

$$\text{Pre}^{\dagger} = \frac{\sum_{p_j \in P} \max_{l_i \in L} \text{Match}(p_j, l_i)}{|P|} \tag{8}$$

$$\text{Rec}^{\dagger} = \frac{\sum_{l_i \in L} \max_{p_j \in P} \text{Match}(p_j, l_i)}{|L|} \tag{9}$$

To make the computation process more intuitive, we illustrate the metrics with a concrete example.

Suppose the ground truth annotations are $L = \{l_1, l_2, l_3\}$ and the model predictions are $P = \{p_1, p_2\}$, where the semantic equivalence matrix is given by:

	l_1	l_2	l_3
p_1	1	1	0
p_2	0	0	1

Then, for this sample case, we compute:

$$\begin{aligned} \text{Pre}^{\dagger} &= \frac{\max(1,0,0) + \max(0,1,0)}{2} \\ &= \frac{1+1}{2} = 1, \\ \text{Rec}^{\dagger} &= \frac{\max(1,0) + \max(1,0) + \max(0,1)}{3} \\ &= \frac{1+1+1}{3} \approx 0.67 \end{aligned} \tag{10}$$

C Error Rate Analysis

We evaluated 109 anomalous cases within the benchmark, of which 21 were misclassified. A detailed quantitative analysis of the error rates at each stage of the CSIAD pipeline was performed for these misclassified cases, as presented in the Table 7.

The failure patterns associated with each step are as follows:

- **Retrieval Ensemble:** Failed to retrieve associated images.
- **Structured Text Extraction:** Incorrect text information extraction.

Table 7: Error statistics at each stage of the CSIAD.

Step	Error Statistics	Error Distribution (%)
1. Retrieval Ensemble	3	14.29
2. Structured Text Extraction	11	52.38
3. Sample-Specific Rule Generation	4	19.05
4. Fact-Driven Rule Verification	0	0.00
5. Rule-Based Anomaly Detection	2	9.52
6. Fact-Driven Anomaly Verification	1	4.76
Total	**21**	**100.00**

- **Sample-Specific Rule Generation:** Failed to generate actionable rules.
- **Fact-Driven Rule Verification:** Misclassification of valid rules as erroneous.
- **Rule-Based Anomaly Detection:** Logical reasoning errors.
- **Fact-Driven Anomaly Verification:** Misjudgment of correctly analyzed anomalies.

We found that the most error-prone stage is **Structured Text Extraction**. Multimodal LLMs may fail to accurately extract textual information due to complex image content (e.g., blurring, deformation, etc.). Incomplete or incorrectly structured text can directly lead to deviations in subsequent steps.

D Our Prompt

D.1 Prompt for Rule Generation

```
Your task is to analyze the provided
    structured text of images and
    identify the rules that need to be
    followed. Please complete the task
    according to the steps below.

Step 1. Sample Type Identification
Analyze the sample content to identify:
1.Category (e.g., invoice, order, tax
    certificate, etc.);
2.All fields included in the sample;
3.Key identifiers involved in the fields
    (Key identifiers refer to fields
    that uniquely identify an object or
    sample, such as ID numbers, contract
     numbers, or user IDs).

Step 2. Rule Generation
Based on the <sample type>, <all fields
    >, and provided <rule examples>
    obtained in Step 1, explore as many
    relevant rules as possible.
    Requirements:
1.Rules must be directly related to the
    sample category. For example,
    invoice rules should relate to
    invoice content, and contract rules
    should relate to contract content.
2.Each rule should use strong constraint
     language (e.g., "must," "should be
    identical"), avoiding ambiguous
    terms like "can" or "usually."
3.Each rule should focus on "information
     consistency" or "logical conflict
    resolution."

Output Format:
<Rule_Generation>
1.Rule xxx
2.Rule yyy
</Rule_Generation>

---
The provided rule examples are as
    follows:
[{rule_list}]

The structured text information of the
    certificate is as follows:
[{text_KVs}]
```

D.2 Prompt for Fact-Driven Rule Verification

```
You are required to analyze the given
    structured text step by step to
    verify whether the provided rule is
    violated within the structured data.
     The analysis process must be fully
    documented before drawing a
    conclusion.

Step 1: Decompose the rule and extract
    the relevant fields.

Step 2: Focus on the structured text and
     extract the fields identified in
    Step 1.

Step 3: Analyze whether the structured
    text extracted in Step 2 aligns with
     the described rule.
- If the data conforms to the rule, the
    rule should be retained.
- Otherwise, the rule should be
    discarded.

Step 4: Output based on the analysis
    process.
Use the following format:
<Rule_Verify>
Retain/Discard
</Rule_Verify>
```

```
---
The reference structured text is as
    follows:
[{fact_case}]

The rule to be verified is as follows:
[{rule}]
```

D.3 Prompt for Anomaly Conclusion Generation

```
You are required to perform cross-sample
     anomaly analysis on multiple
    transaction samples and identify
    potential anomalies. Please strictly
     follow the requirements below to
    complete the task.

Step 1: Analyze the fields mentioned in
    the provided rule.

Step 2: Analyze the structured text of
    the sample group and extract the
    content related to the fields
    mentioned in Step 1.

Step 3: Verify whether there are any sub
    -sample groups violate the given
    rule. If violations are found,
    indicate the affected record IDs and
     analyze the cause of the anomaly.

Step 4: Output the result in the
    following format:
- For anomalies: If anomalies exist,
    output in the following format:
<Anomaly>
Violation of rule<Rule Description>,
Affected records: [<Record ID1>, <Record
     ID2>, ...],
Explanation: <Anomaly Explanation>
</Anomaly>

- For no anomalies: If no anomalies are
    found, output the following:
<Anomaly>
No anomaly
</Anomaly>

---
The rules to be verified are as follows:
[{rule_list}]

The structured text information of the
    sample group to be analyzed is as
    follows:
[{text_KVs}]
```

D.4 Prompt for Fact-Driven Anomaly Verification

Judge if the anomaly conclusion is consistent with the structured text of the certificate image:

```
Step1: Extract from Explanation
Extract the <fields> and <field content>
     mentioned in the explanation from
    the Anomaly Conclusion;

Step2: Extract from RawText
Extract the field content related to the
     mentioned fields in Step1 from the
    Structured Texts.

Step3:Combine the Field Content,
Consolidate the content extracted in
    Step 1 and Step 2 for the same
    fields within the same record ID
    into a unified list.

---
The given anomaly conclusion:
[{anomaly_conclusion}]
Structured text information of the image
     group: [{text_KVs}]
```

After LLM integration, use a regular expression to check whether the field content mentioned in the anomaly conclusion is consistent with the original text.

Judge if the anomaly conclusion aligns with the rule description:

```
Your task is to proofread the cross-
    sample anomaly conclusions for
    multiple transaction samples and
    verify the authenticity of potential
     anomalies. Please analyze step by
    step.

Step 1: Anomaly Confirmation
Analyze the following content:
- Mentioned anomaly: Violation of the
    rule <Rule Description>
- Affected records: <Record ID1>, <
    Record ID2>
- Conclusion 1: Does it satisfy the rule
    's premise: Yes/No
- Conclusion 2: Does it violate the rule
    : Yes/No
- Decision result: Anomaly confirmed/
    Anomaly not confirmed
// The anomaly is considered confirmed
    if both conclusions are "Yes." If at
     least one is "No," the anomaly is
    considered not confirmed.

Step 2: Anomaly Conclusion Revision
Based on the analysis results in Step 1,
     revise the anomaly conclusion and
    output the revised conclusion in the
     following format:
<Conclusion_Verify>
Anomaly confirmed/not confirmed
</Conclusion_Verify>

---
The given anomaly conclusion:
[{anomaly_conclusion}]
```

E Case Study

E.1 Complete Workflow of Cross-Sample Analysis.

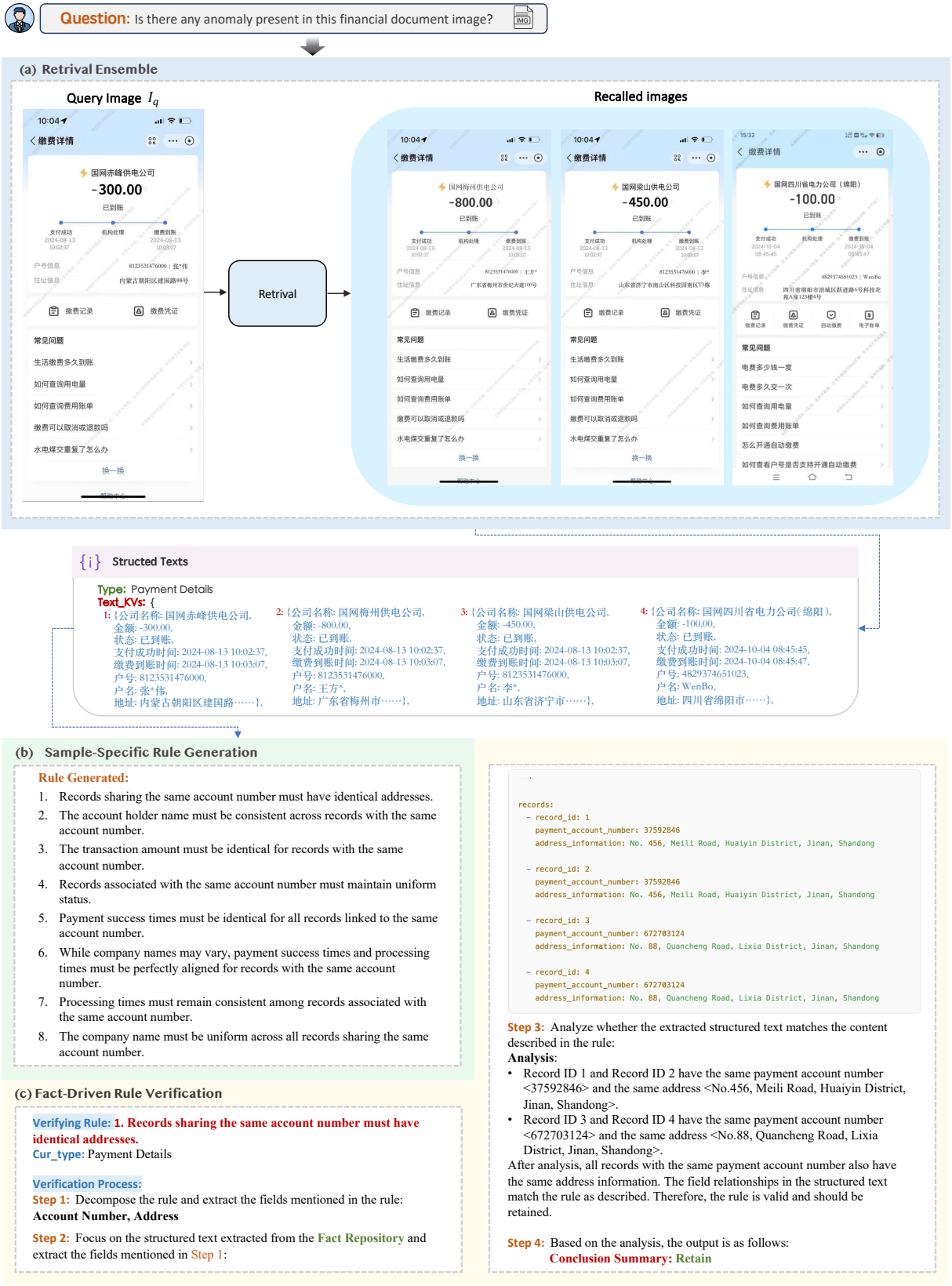

Figure 5: Complete Workflow of Cross-Sample Analysis (1/3).

Verifying Rule: **2.The account holder name must be consistent across records with the same account number**
Cur_type: Payment Details

Verification Process:
Step 1: Decompose the rule and extract the fields mentioned in the rule:
Account Number, Account Holder Name

Step 2: Focus on the structured text extracted from the **Fact Repository** and

```
records:
  - record_id: 1
    account_holder_name: Jinan Huaiyin East and Hardware Store
    payment_account_number: 37592846

  - record_id: 2
    account_holder_name: Jinan Huaiyin East and Hardware Store
    payment_account_number: 37592846

  - record_id: 3
    account_holder_name: Jinan Huaiyin Youxi Dry Cleaning Store
    payment_account_number: 672703124

  - record_id: 4
    account_holder_name: Jinan Huaiyin Youxi Dry Cleaning Store
    payment_account_number: 672703124
```

Step 3: Analyze whether the extracted structured text matches the content described in the rule:
Analysis:
- Record ID 1 and Record ID 2 share the same payment account number (37592846), and the account holder name is Jinan Huaiyin East and Hardware Store for both.
- Record ID 3 and Record ID 4 share the same payment account number (672703124), and the account holder name is Jinan Huaiyin Youxi Dry Cleaning Store for both.

After analysis, all records with the same account number also have consistent account holder names. The relationships between fields in the structured text match the rule as described. Therefore, the rule is valid and should be retained.

Step 4: Based on the analysis, the output is as follows:
Conclusion Summary: Retain

Verifying Rule: **3.The transaction amount must be identical for records with the same account number**
Cur_type: Payment Details

Verification Process:
Step 1: Decompose the rule and extract the fields mentioned in the rule:
Account Number, Amount

Step 2: Focus on the structured text extracted from the **Fact Repository** and

```
records:
  - record_id: 1
    payment_account_number: 37592846
    amount: 10.00

  - record_id: 2
    payment_account_number: 37592846
    amount: 10.00

  - record_id: 3
    payment_account_number: 672703124
    amount: 20.00

  - record_id: 4
    payment_account_number: 672703124
    amount: 23.00
```

Step 3: Analyze whether the extracted structured text matches the content described in the rule:
Analysis:
- For Payment Account Number 37592846, both Record ID 1 and Record ID 2 have the same amount of 10.00.
- For Payment Account Number 672703124, Record ID 3 has an amount of 20.00, while Record ID 4 has an amount of 23.00.

After analysis, the samples with Payment Account Number <37592846> have consistent amounts, but the samples with Payment Account Number <672703124> have inconsistent amounts. Therefore, the relationships between the fields in the structured text do not align with the rule.

Step 4: Based on the analysis, the output is as follows:
Conclusion Summary: Discard

Rules 4-8 follow the same rule verification steps. After verification by the fact repository, the rules that are consistent with the patterns observed in normal samples. The following are all the rules that have been retained:

Rule Verified:

1. Records sharing the same account number must have identical addresses.
2. The account holder name must be consistent across records with the same account number.
4. Records associated with the same account number must maintain uniform status.
8. The company name must be uniform across all records sharing the same account number.

(d) Rule-Based Anomaly Detection

Analysing Rule: **1.Records sharing the same account number must have identical addresses**

Analysis Process:
Step 1: Rule Analysis
The rule states that if multiple electricity bill records share the same account number, the addresses for these samples must be identical. The relevant fields are **<Account Number>** and **<Address>.**

Step 2: Analysis of given Structured Texts
The content relevant to the above fields is extracted as follows:

```
records:
  - record_id: 1
    account_number: 8123531476000
    address: "No. 88, Jianguo Road, Chaoyang District, Inner Mongolia"

  - record_id: 2
    account_number: 8123531476000
    address: "No. 100, Shiji Avenue, Meizhou City, Guangdong Province"

  - record_id: 3
    account_number: 8123531476000
    address: "Building T3, Southern Section, Science and Technology Park, Nanshan Distric

  - record_id: 4
    account_number: 4829374651023
    address: "No. 4, Floor 123, Block A, Kejihuayuan, No. 6 Yuejin Road, Fucheng District
```

Step 3: Verification of Rule Violation
Records 1,2 and 3 share the same account number, <8123531476000>, but the addresses differ.

Step 4: Analysis Result Output
- **Conclusion:** Violation of rule "Records sharing the same account number must have identical addresses".
- **Affected records:** [1, 2, 3]
- **Explanation:** All account numbers are < 8123531476000 >, but the addresses are <No.88 Jianguo Road…>, <No.100 Shiji Avenue…>, <Building T3, Southern Section…>, respectively.

Conclusion Summary: Anomaly

(e) Fact-Driven Anomaly Verification

Verifying Anomaly: Violation of the **rule** *<If multiple electricity bill records share the same account number, the addresses for these samples must be identical>*. **Affected records**: [1, 2, 3]. **Explanation**: All account numbers are < 8123531476000 >, but the addresses are <No.88 Jianguo Road…>, <No.100 Shiji Avenue…>, <Building T3, Southern Section…>, respectively.

Verification Process:
Step 1: Confirm whether the fields are faithful to the original text
Fields by explanation:

```
records:
  - record_id: 1
    account_number: 8123531476000
    address: "No. 88, Jianguo Road, Chaoyang District, Inner Mongolia"

  - record_id: 2
    account_number: 8123531476000
    address: "No. 100, Shiji Avenue, Meizhou City, Guangdong Province"

  - record_id: 3
    account_number: 8123531476000
    address: "Building T3, Southern Section, Science and Technology Park, Nanshan Distric
```

Figure 6: Complete Workflow of Cross-Sample Analysis (2/3).

Fields by raw text:

```
records:
  - record_id: 1
    account_number: 8123531476000
    address: "No. 88, Jianguo Road, Chaoyang District, Inner Mongolia"

  - record_id: 2
    account_number: 8123531476000
    address: "No. 100, Shiji Avenue, Meizhou City, Guangdong Province"

  - record_id: 3
    account_number: 8123531476000
    address: "Building T3, Southern Section, Science and Technology Park, Nanshan Distric
```

Fields combined:

```
fields:
  - record_id: 1
    account_number:
      - "8123531476000"
      - "8123531476000"
    address:
      - "No. 88, Jianguo Road, Chaoyang District, Inner Mongolia"
      - "No. 88, Jianguo Road, Chaoyang District, Inner Mongolia"

  - record_id: 2
    account_number:
      - "8123531476000"
      - "8123531476000"
    address:
      - "No. 100, Shiji Avenue, Meizhou City, Guangdong Province"
      - "No. 100, Shiji Avenue, Meizhou City, Guangdong Province"

  - record_id: 3
    account_number:
      - "8123531476000"
      - "8123531476000"
    address:
      - "Building T3, Southern Section, Science and Technology Park, Nanshan District, Ji
      - "Building T3, Southern Section, Science and Technology Park, Nanshan District, Ji
```

Field Confirmation Conclusion: The fields match the original text

Step 2: Confirm whether the mentioned anomaly matches the rule

- **Conclusion 1:** Does the rule's premise hold?

The premise of the rule is "*If multiple electricity bill records share the same account number*", and the "account number" field in all affected records is < 8123531476000 >, which is consistent across all records. Therefore, the premise is met, and Conclusion 1 is "**Yes**".

- **Conclusion 2:** Does the rule's conclusion hold?

The conclusion of the rule is "*The addresses must be identical*", but the addresses for the affected records are clearly different, so they do not satisfy the rule's conclusion, resulting in a violation. Conclusion 2 is "**Yes**".

Anomaly Confirmation Conclusion: The mentioned anomaly matches the rule

Decision: Anomaly confirmed

The anomaly analysis processes for **Rules 2, 4**, and **8** are the same as the anomaly analysis process for Rule 1 mentioned above, and are therefore omitted here. The final conclusion is that the financial document images to be inspected exhibit multiple anomalies, as detailed below:

For Rule 2: Violation of rule "*The account holder name must be consistent across records with the same account number*".

- **Affected records:** [1,2,3]
- **Explanation:** All account numbers are <8123531476000>, but the account holder names are <Zhang*Wei>, <WangFang*>, <Li*>.

For Rule 4: No violations were found for the rule: "*Records associated with the same account number must maintain a consistent status*".

- **Affected records:** None
- **Explanation:** None

For Rule 8: Violation of rule "*The company name must be uniform across all records sharing the same account number*".

- **Affected records:** [1,2,3]
- **Explanation:** All account numbers are <8123531476000>, but the company names are <State Grid Chifeng Power Supply Company>, <State Grid Meizhou Power Supply Company>, <State Grid Liangshan Power Supply Company>.

Figure 7: Complete Workflow of Cross-Sample Analysis (3/3).

E.2 Comparison Case

Figure 8: A comparison case for Fact-Based Rule Verification and LLM Self-Reflection-Based Rule Verification. The **Left** column illustrates the verification process of a rule that should be retained, while the **Right** column depicts the process for a rule that should be discarded. As shown, the conclusions drawn from Fact-Based Rule Verification are accurate in both cases, whereas the conclusions produced by the LLM Self-Reflection-Based approach are incorrect.

F Dataset Composition

Table 8: The distribution of anomalous certificate images in the evaluation data. The *Type* includes the financial document image type and the description of the anomaly type. *Group Num.* refers to the number of financial document image groups of this type, while *Image Num.* refers to the total number of financial document images contained within these groups.

Type	Group Num.	Image Num.
Tax Payment Certificate_same certificate number_different internal information_2	8	26
Payment Details_same user ID_different other user information_1	7	21
Tax Payment Certificate_same certificate number_different internal information_1	7	22
Tax Payment Certificate_same certificate number_different internal information_3	6	28
Mini Program Info_same Mini Program ID_different Mini Program name_1	6	20
No Tax Arrears Certificate_same certificate number_different internal information_1	6	19
Handheld Commitment Letter_same identity information_multiple registered email accounts_1	5	23
Taobao Store Operation Certificate_same store address_different internal information_1	5	28
No Tax Arrears Certificate_same certificate number_different internal information_2	3	9
Payment Details_same user ID_different other user information_3	3	14
Life Payment Voucher_same household number_different address information_1	3	11
Tax Payment Certificate_same certificate number_different internal information_4	2	11
Tax Certificate_same certificate number_different internal information_4	2	12
Tax Payment Certificate_same taxpayer identification number_different taxpayer name_1	2	8
Tax Payment Certificate_same verification code_different internal information_1	2	4
No Tax Arrears Certificate_same certificate number_different internal information_3	2	7
Sales Contract_same contract number_different internal information_1	2	6
Permanent Resident Registration Card_same ID number_different name_1	2	4
Tax Certificate_same certificate number_different internal information_2	1	3
VAT Electronic General Invoice_same invoice number_different internal information_1	1	5
Authorization Letter_same authorization dealer contract number_different authorized person_1	1	5
Marriage Certificate_same marriage certificate number_different internal information_1	1	3
House Lease Contract_same lessee (Party B) ID number_different name_1	1	2
Payment Details_same user ID_different other user information_2	1	6
House Lease Contract_same lessor (Party A) ID number_different name_1	1	2
Sales Contract_same contract number_different internal information_2	1	4
Commercial Office Lease Contract_same lessor (Party A) ID number_different name_1	1	4
ID Card Front_same ID number_different name_1	1	2
Payment Details_same user ID_different other user information_4	1	6
Sales Contract_same contract number_different agent_1	1	2
Business License_same unified social credit code_different business owner_1	1	2
Purchase and Sales Contract_same order number_different internal information_1	1	2
Tax Certificate_same certificate number_different internal information_3	1	3
JD Order Details_same order number_different internal information_1	1	2
Tax Certificate_same certificate number_different internal information_1	1	2
Tax Payment Certificate_same certificate number_different taxpayer name_1	1	2
Promissory Note_same lender's ID number_different name_1	1	4
Promissory Note_same lender's ID number_different name_2	1	6
Promissory Note_same lender's ID number_different name_3	1	2
Outbound Order_same order number_different internal information_1	1	2
Proof of Employment_same employee ID number_different name_1	1	4
VAT Electronic General Invoice_same invoice number_different internal information_1	1	2
Authorization Letter_same authorization dealer contract number_different authorized person_1	1	3
Income Tax Details_same taxpayer identification number_different taxpayer name_1	1	6
No Tax Arrears Certificate_same taxpayer identification number_different issuing unit_1	1	3
Taobao Store Operation Certificate_same store address_different internal information_2	1	6
Taobao Store Operation Certificate_same store address_different internal information_3	1	3
Tax Payment Certificate_same taxpayer identification number_different taxpayer name_2	1	2
Promissory Note_same borrower ID number_different name_1	1	2
Tax Payment Certificate_same certificate number_different internal information_5	1	6
Tax Payment Certificate_same certificate number_different internal information_6	1	4
Tax Payment Certificate_same certificate number_different internal information_7	1	3
Tax Payment Certificate_same certificate number_different issuance date_1	1	2
Tax Certificate_same certificate number_different internal information_4	1	6
Total	109	396

Table 9: The distribution of non-anomalous financial document images in the evaluation data. The *Type* includes only the description of the financial document image type. *Group Num.* refers to the number of financial document image groups of this type, while *Image Num.* refers to the total number of financial document images within these groups. The *Group Num.* and *Image Num.* are the same because non-anomalous images do not belong to associated image groups, and each non-anomalous image is treated as an individual.

Type	**Group Num.**	**Image Num.**
National Emblem Side of ID Card	14	14
Account Details	9	9
Business License	7	7
Bill Details	7	7
ID Card Front Side	6	6
Income and Expenditure Record	6	6
Taobao Store Operation Certificate	5	5
Detail Inquiry	5	5
Transaction Details	4	4
Payment Details	4	4
Bank Card	3	3
No Tax Arrears Certificate	3	3
Passport	3	3
Courier Receipt	3	3
Foreign Card Details	3	3
Income and Expenditure Details	2	2
Income and Expenditure Information	2	2
Permanent Resident Registration Card	2	2
Tax Payment Certificate	2	2
Order Details	2	2
Transfer Record	1	1
Account Details	1	1
Repayment Details	1	1
Pass	1	1
Wallet Records	1	1
Transaction Details	1	1
Contact Person	1	1
Logistics Details	1	1
Repayment Inquiry Details	1	1
Inquiry Details	1	1
Transaction Flow	1	1
Record Inquiry	1	1
Medical Form	1	1
Sale Listing	1	1
Promissory Note	1	1
Balance Details	1	1
Detail Information	1	1
Total	109	109

G Human Annotation Cases

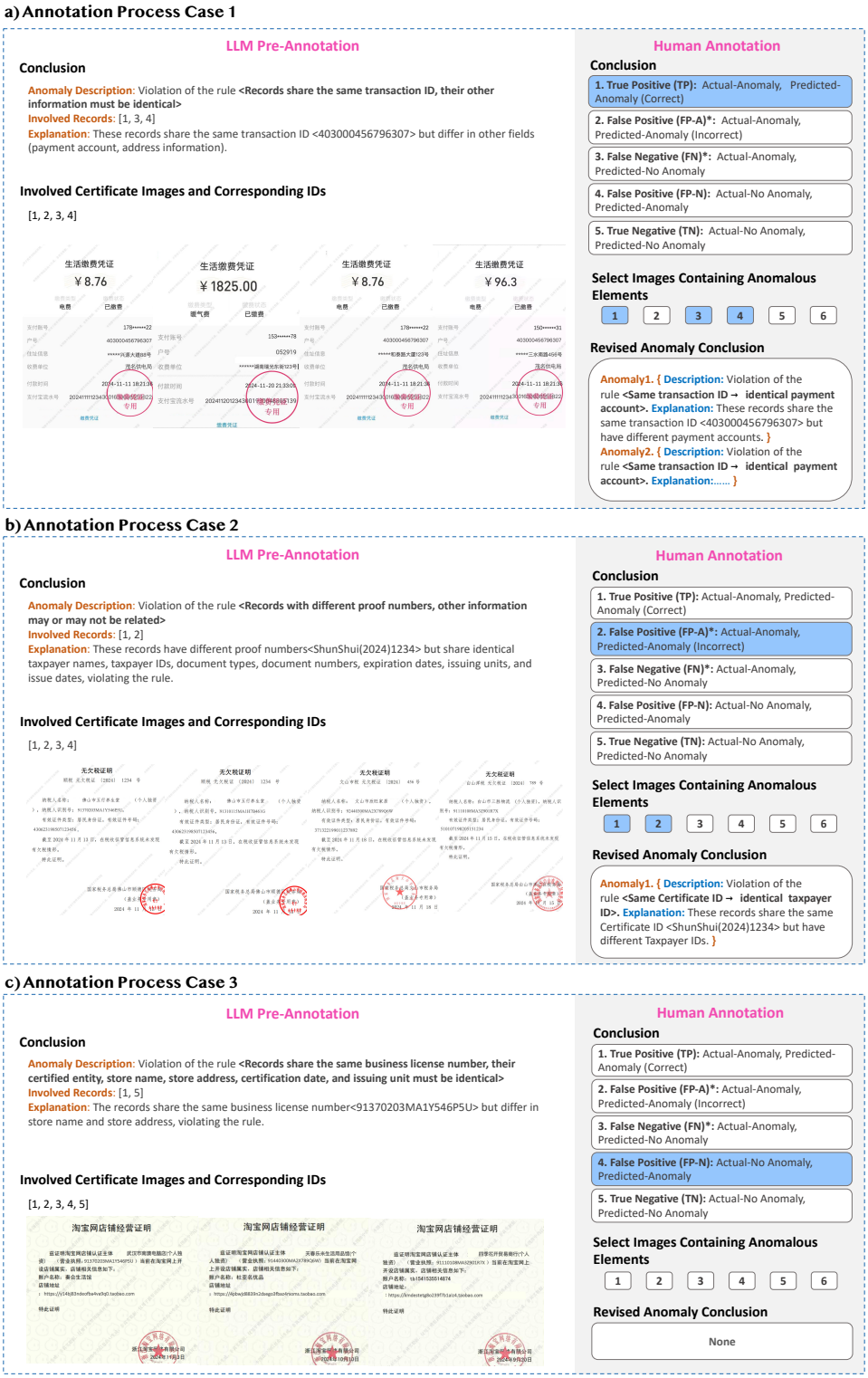

Figure 9: Examples of the Human Annotation Process. Given the LLM's pre-annotated conclusions, human annotators need to determine whether anomalies exist in the actual images and whether the LLM's pre-annotated conclusions align with the ground truth. They then select one of five predefined labels (TP, FP-A, FN, FP-N, TN) to evaluate the LLM's pre-annotated results. If the selected label is FP-A or FN (marked with *), it indicates that the LLM failed to identify the correct anomalous elements, and annotators are required to correct or supplement the anomalous elements along with the associated images. Figures (a), (b), and (c) demonstrate examples of LLM pre-annotated results as TP, FP-A, and FN-N, respectively.