CogniCred: A Dataset and Benchmark for Cognitive Credential Forgery Detection

Junchi Li[1*], Jiasheng Sun[1*], Weizhi Chen[1], Ziwei Wang[1], Sheng Zhou[2†], Jiajun Bu[2], Chenfan Qu[3], Bohan Yu[4], Jian Liu[4†], and Weiqiang Wang[4]

[1] College of Computer Science and Technology, Zhejiang University, China
[2] Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, Zhejiang University, China
[3] South China University of Technology
[4] Ant Group
{junchili,22521132,wangziwei98, zhousheng_zju, chenweizhi, bjj}@zju.edu.cn, rex.lj@antgroup.com

Abstract. Credentials, such as corporate qualifications and personal identification, serve as common forms of evidentiary materials. The forgery of such credentials poses significant risks to security, making its detection a long-standing area of widespread concern. With the advancement of AI-Generated Content technologies, forgery techniques can now largely preserve visual characteristics, making purely vision-based detection methods insufficient to meet practical demands. Instead, a growing number of forgeries necessitate cognitive reasoning for detection. While the development of Multimodal Large Language Models (MLLMs) presents new opportunities for complex credential forgery detection, the scarcity of datasets and benchmarks severely impedes progress in this critical domain. To bridge this gap, we introduce CogniCred — the first large-scale, multimodal dataset built upon complex, real-world credentials with carefully crafted cognitive-level forgeries. Along with this dataset, we propose a dedicated benchmark, namely CogniCredBench, that evaluates model performance across five distinct types of cognitive forgeries. We conduct comprehensive evaluations across 21 mainstream MLLMs, revealing substantial performance disparities and a strong bias toward text-based reasoning, highlighting that current models still fall short in balanced multimodal reasoning for cognitive forgery detection. Our datasets are publicly available at `https://huggingface.co/datasets/sjs0606/CogniCred`.

Keywords: Document Forgery Detection · Multimodal Large Language Models · Visual Reasoning

1 Introduction

Credentials, such as purchase receipts, ID cards, and chat records, are critical in a wide range of real-world, security-sensitive applications. As these creden-

[*] Equal contribution.
[†] Corresponding author.

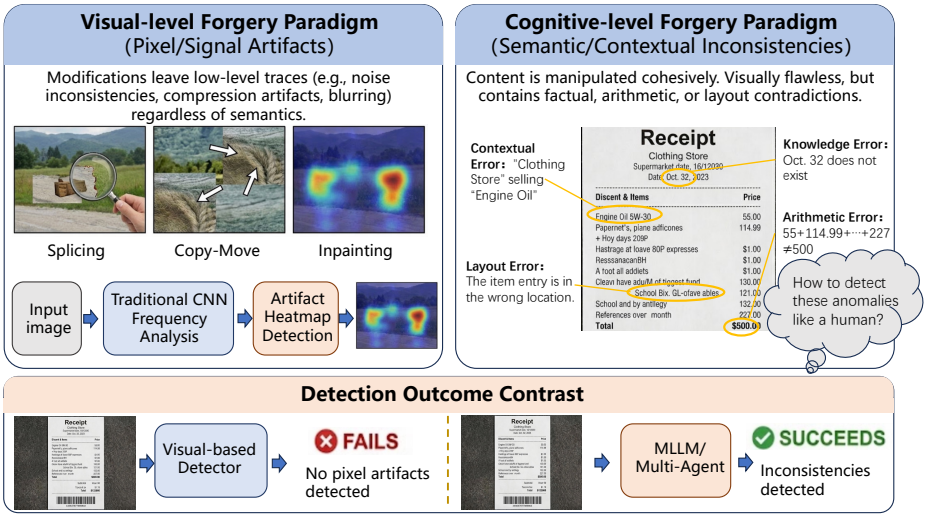

Fig. 1: Illustration of detecting different forgeries, where Cognitive-level forgeries cannot be detected by methods relying solely on visual cues.

tials increasingly become targets of forgery, a variety of techniques have emerged to tamper with them in both visible and covert ways. Traditional visual forgery methods including copy-paste operations, pixel-level inpainting, or content splicing—typically leave behind detectable artifacts. These can be detected through methods that analyze visual cues such as visual appearance inconsistencies [5,21] and texture irregularities [22, 29], with representative techniques including blur artifact analysis [14], frequency coefficient anomaly detection [4], compression anomaly detection [11], and splicing localization [3].

The advancement of generative AI has led to a substantial increase in the sophistication of forgery techniques. Cutting-edge methods are now capable of making credential forgeries that are typically integrated meticulously into intricate layout designs, while maintaining visual authenticity [40], which makes it challenging to notice visual forgeries at a superficial level. As illustrated in Figure 1, beyond visual imperfections, these credentials may exhibit forgeries in cognitive aspects—that is, manipulations that are visually flawless but harbor underlying semantic, structural, or factual contradictions—such as mismatched names, contradictory timestamps, or unrealistic numerical totals, which necessitate cross-modal and semantic reasoning for detection. This underscores an alternative detection strategy: rather than concentrating solely on visual indicators, it requires the models to identify inconsistencies in the fundamental semantics and structure of the credentials.

Currently, existing benchmarks focus primarily on detecting pixel-level visual alterations—such as localized tampering or splicing [20,38] on credentials, while benchmarks for cognitive-level credential forgery detection remain insufficient. Although there are some general multimodal benchmarks that do test for cogni-

tive reasoning, they typically assess a model's grasp of open-world common sense, which is actually different from verifying the internal consistency of a credential against its structured, rule-based constraints. The lack of cognitive-focused datasets and comprehensive benchmarks may significantly hinder the advancement of detection technologies, leaving critical security applications increasingly vulnerable to undetected forgeries in the era of AI-generated content.

To bridge this gap, we introduce CogniCred, the first dataset designed to assess the capability of MLLMs in detecting cognitive-level credential forgeries. CogniCred is built upon over 16,000 meticulously curated samples whose diversity reflects real-world complexity—spanning rigid tabular layouts of financial invoices, semi-structured formats of official certificates, and conversational, flow-based designs of chat transcripts. This dataset is organized into five categories, each representing a distinct reasoning capability for different kinds of forgeries, and every category is further divided into 2-3 fine-grained subtypes. All samples are annotated with labels, including forgery region masks, forgery-type tags, and corresponding reasoning statements, enabling not only model training and evaluation but also interpretability analysis and error diagnosis.

Based on this dataset, we further construct the benchmark around five categories to comprehensively assess the model across five complementary reasoning dimensions. For each dimension, the benchmark supports two evaluation tasks: a closed-set classification task and an open-ended reasoning task, aimed at testing the model's forgery detection capabilities under distinct conditions.

We evaluate 21 state-of-the-art MLLMs—including 18 open-source and 3 closed-source models—along with human performance on CogniCred. Results show that CogniCred remains highly challenging: model accuracy on visually grounded reasoning tasks is generally only 20%–30%, and models of the same family tend to exhibit similar performance, regardless of model scale. Our main contributions are as follows:

- We introduce **CogniCred**, a novel dataset and benchmark for the detection of cognitive forgery in credentials. It includes 5 major forgery types and 14 subcategories, accompanied by a structured evaluation protocol designed to probe MLLMs' cognitive reasoning capabilities.
- We conduct systematic benchmarking of mainstream MLLMs on CogniCred, revealing significant performance gaps between proprietary and open-source models and demonstrating the benchmark's effectiveness in exposing deficiencies in cognitive multimodal reasoning.
- We observe a critical failure mode in current MLLMs: an over-reliance on textual cues while ignoring visual and layout inconsistencies, revealing a fundamental gap in their multimodal understanding and vision grounding.

2 Related Work

2.1 Credential Forgery Detection

This work is situated in the domain of credential forgery detection, which aims to identify unauthorized alterations in credential images. Early approaches focused

on detecting tampering in idealized document images using rule-based methods, such as verifying the alignment of fonts [2] or text lines [34]. However, these rule-based approaches perform poorly on casually photographed credentials of various types, as they rely heavily on manually designed rules [30]. To address this limitation, several benchmarks with casually captured credential images have been recently introduced [16, 20, 38]. Recent approaches have primarily targeted visual-level inconsistencies—such as anomalies in visual appearance [5, 21], frequency coefficients [4], and texture patterns [22, 29]. These methods [19, 23, 26], however, tend to overlook artifacts related to semantics and logic, which are key indicators for forgery detection in real-world scenarios. Other lines of work focus on specific forensic cues in particular credentials, such as face-swapping detection [18], digital watermark verification, or signature authentication. Nevertheless, these approaches are typically limited to a single modality and lack systematic cross-element modeling.

2.2 Multimodal Forgery Reasoning

Recent datasets for evaluating models' forgery detection capabilities have increasingly incorporated multimodal forgery reasoning, with particular emphasis on content-consistency detection—a fundamental component of cross-modal reasoning. This research direction originated in the NLP domain, exemplified by tasks such as fact-checking [33] and summary inconsistency detection [12], and has subsequently been extended to vision–language settings. For instance, MMIR [41] introduces image-text consistency tasks by injecting semantic anomalies (e.g., identity or temporal conflicts) into synthetic visual documents. However, its relatively small scale (500 samples) constrains the model's ability to generalize. MFC-Bench [36] offers a larger dataset for detecting factual inconsistencies in image–text pairs but focuses mainly on shallow visual-level reasoning. MMDOCBench [45] advances document understanding via multimodal QA and information extraction, improving layout perception yet lacking coverage of structured credentials and logic-based anomalies. IDNet [7] addresses identity forgery with field-level annotations and image variants but omits unified reasoning tasks and cross-document generalization.

Another key research direction is AI-generated content detection. ForensicsBench [35] expands to 63K samples across multiple modalities (including GANs, diffusion models, and video), yet remains limited to binary authenticity classification without structural or semantic reasoning. DD-VQA [44] and FFAA [10] evaluate large models in deepfake detection, but their synthetic samples emphasize visual level artifacts or manipulation traces rather than logical inconsistencies. A growing body of work—including [8], FakeBench [13], LOKI [43], TextShield-R1 [25] and MIML [24]—has explored the potential of large multimodal models (LMMs) in synthetic detection, highlighting their ability to provide natural language explanations for true/false predictions, thereby enhancing interpretability. In a different vein, [39] targets logical contradictions in natural scenes via common sense, this approach is insufficient for the strict, rule-based logical constraints of official credentials.

Despite progress in vision-language alignment, authenticity classification, and layout parsing, no existing benchmark systematically evaluates the semantic reasoning capabilities of MLLMs in structured credential scenarios—a key gap our work aims to address.

3 CogniCred

3.1 Overview

CogniCred introduces a new perspective on forgery detection by constructing credential samples with cognitive anomalies, shifting the evaluation focus from traditional visual artifact detection to assessing a model's cognitive reasoning capability. To systematically characterize these cognitive-level forgeries, we propose a hierarchical taxonomy encompassing five mutually independent reasoning dimensions: *Internal Reasoning*, *World Knowledge*, *Visual Reasoning*, *Contextual Reasoning*, and *Format Compliance*. These dimensions are further decomposed into 14 fine-grained sub-categories representing distinct, real-world credential forgery scenarios, as illustrated in Figure 2.

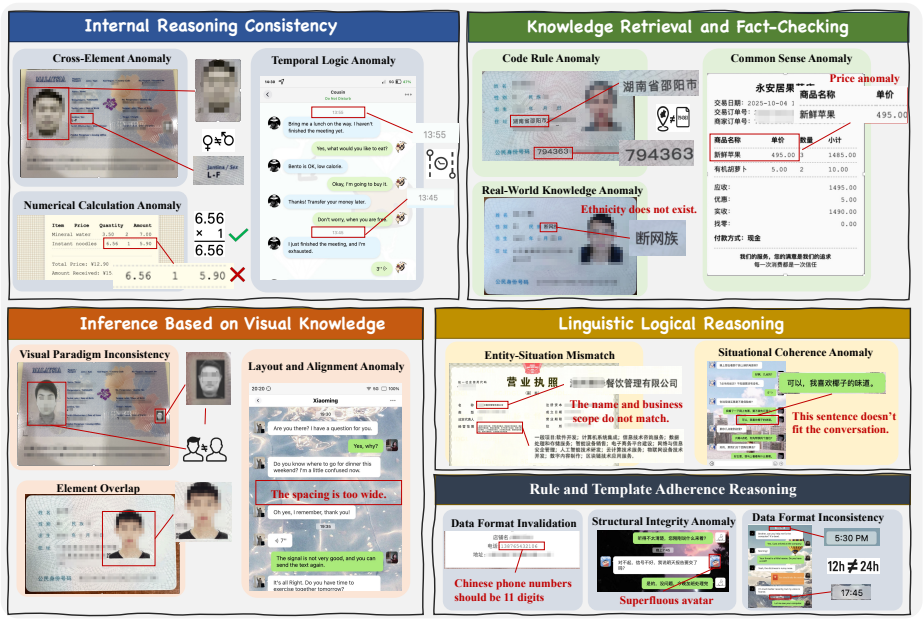

Fig. 2: Taxonomy of CogniCred dataset. The red markings precisely indicate the cognitive anomalies in each example, emphasizing the dataset's focus on complex semantic contradictions rather than pixel-level artifacts.

Internal Reasoning–*Internal Reasoning Consistency*. This category assesses logical coherence within a closed system of credential elements, encompassing

three fine-grained sub-categories: *(a) Numerical Calculation forgery*, which denotes arithmetic inconsistencies in computed fields (*e.g.*, conflicting totals on shopping receipts); *(b) Temporal Logic forgery*, referring to violations of chronological rationality in time-related information (*e.g.*, anomaly in the chronological order of chat messages); and *(c) Cross-Element forgery*, indicating mismatches between interdependent fields (*e.g.*, the stated gender or age is inconsistent with the appearance of the person in the ID photo).

World Knowledge–*Knowledge Retrieval and Fact-Checking.* This category evaluates the model's capability to identify implausible or factually erroneous content by harnessing external knowledge, encompassing three sub-categories: *(a) Code Rule forgery*, which checks misalignment with official coding schemes (*e.g.*, mismatch between administrative codes and address information); *(b) Common Sense and Reasonableness forgery*, which flags unreasonable values (*e.g.*, exorbitant item prices on receipts or malformed time formats in chat logs); and *(c) Real-World Knowledge forgery*, which identifies non-existent entities (*e.g.*, invalid ethnicities in identity credentials).

Visual Reasoning–*Inference Based on Visual Knowledge.* This category centers on inconsistencies stemming from visual layout, appearance, or multimodal alignment, encompassing three sub-categories: *(a) Visual Paradigm Inconsistency*, which refers to violations of expected visual identity norms, (*e.g.*, mismatched photos in international passports); *(b) Layout and Alignment forgery*, which denotes irregular spatial arrangements, (*e.g.*, abnormal spacing between chat bubbles); and *(c) Element Overlap or Out of Bounds*, which captures cases where content is misplaced, overlapping, or rendered beyond its designated region, (*e.g.* text exceeding bubble boundaries in chat logs).

Contextual Reasoning–*Situational Linguistic Logical Reasoning.* This category assesses semantic coherence and contextual plausibility in natural language interactions, encompassing two sub-categories: *(a) Situational Coherence forgery*, which detects semantically disjointed exchanges, (*e.g.*, incoherent question–answer pairs in chat logs); and *(b)Entity-Situation Mismatch*, which identifies implausible combinations, (*e.g.*, business scopes in licenses that are irrelevant).

Format Compliance–*Rule and Template Adherence Reasoning.* This category assesses strict adherence to standardized data formats and structural templates, encompassing three sub-categories: *Data Format Invalidation*, which checks violations of field-specific encoding rules (*e.g.*, overlong phone numbers); *Structural Integrity forgery*, which ensures all required fields are present and no extraneous elements appear (*e.g.*, missing entries on ID cards); and *Data Format Inconsistency*, which enforces uniform formatting across related fields, (*e.g.* inconsistent timestamp styles in chat logs).

Leveraging the aforementioned taxonomy, we propose a four-stage data construction pipeline (see Figure 3) for the automatic generation of large-scale, high-quality forged credential images tailored to each sub-category. The dataset covers five major types of credentials commonly found in real-world scenarios: *ID cards*, *passports*, *business licenses*, *chat transcripts*, and *shopping receipts*.

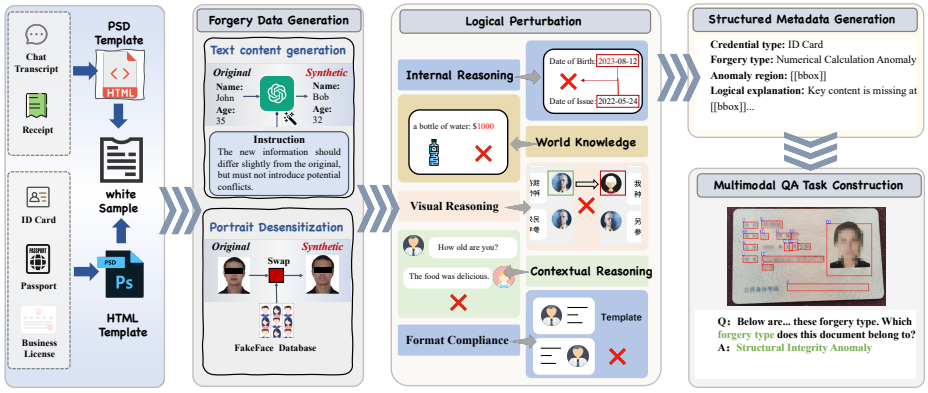

Fig. 3: Construction Pipeline of CogniCred. The pipeline consists of four stages: template construction, content injection, cognitive anomaly injection, followed by structured annotation generation and multimodal QA task formulation.

3.2 Dataset Construction Pipeline

Template Construction. To support high-fidelity forgery generation across diverse credentials, we categorize credential types into two groups—*layout-flexible* and *layout-fixed*—with dedicated generation pipelines tailored to each.

HTML-based Templates. For layout-flexible credentials such as chat transcripts and shopping receipts, we employ structured HTML templates inspired by mainstream UI designs and real-world receipt formats. Instead of conventional placeholder replacement, we employ a semantic field-aware rendering mechanism: it parses structured textual inputs (*e.g.*, timestamps, speaker identity), and dynamically renders visual elements based on field-level semantics. This rendering mechanism supports flexible layout composition, high visual diversity, and precise field positioning, enabling automatic content injection without manual intervention.

PSD-based Templates. For layout-fixed credentials such as ID cards, passports, and business licenses, we construct editable Photoshop (PSD) templates, where each semantic field is defined as a labeled layer with spatial metadata. Based on 100 real-world samples per credential type, we manually design templates to faithfully reproduce authentic visual layouts. Programmatic layer substitution enables scalable synthesis of visually realistic forgeries while preserving layout fidelity and consistency.

Content Injection. We construct visually realistic and semantically consistent credential samples by injecting GPT-4o–generated field values into predefined HTML or PSD templates. For each credential type, we build representative exemplars to guide GPT-4o in generating semantically and format-compliant content, automatically filtering out invalid outputs via predefined rules.

For *HTML-based templates*, field values are rendered into DOM components (*e.g.*, timestamps, chat bubbles) based on semantic mappings. Pyppeteer is used to automate rendering, screenshot capture, and extraction of field-level bounding

boxes. To enhance visual diversity, we further prompt GPT-4o to generate varied CSS themes and crawl over 4,000 background and avatar images.

For *PSD-based templates*, the generated field values are injected into structured JSON files that define the content of editable layers. We use Photoshop JSX scripts to batch-update text layers and export rendered images. For fields involving facial portraits, we apply gender-matched face replacement using the Generated Photos Academic Dataset, achieving anonymization while preserving pose, lighting, and visual realism. The synthesized credential images contain entirely artificial content with no real individuals, thus avoiding ethical and privacy concerns.

Cognitive Anomaly Injection. Based on the semantically consistent credential images obtained in the previous stage, we further construct cognitively anomalous samples by introducing minimal semantic-level perturbations to key fields, thereby injecting cognitive-level forgeries.

Forgeries are introduced based on a set of predefined cross-field semantic consistency constraints (i.e., dependency mappings), such as the alignment between gender and facial appearance. Given a pair of semantically dependent fields (A_1, A_2), we sample a counterfactual but valid value A_3 from the value space of A_1, constructing an inconsistent yet syntactically valid field pair (A_3, A_2).

For HTML-based templates, we use Pyppeteer to update target DOM fields and re-render screenshots; for PSD-based templates, we inject perturbed values into linked JSON configs and invoke a predefined JSX script to auto-redraw the relevant layers.

Post-Processing and Dataset Organization. We ensure data quality via the following procedures: ***(i) Metadata annotation.*** Each sample is annotated with bounding boxes, field mappings, semantic dependency links, and forgery labels. We then derive QA items (e.g., multiple-choice questions) from the specific cognitive anomalies present in the sample. ***(ii) Quality control.*** Human reviewers filter out samples that contain unintended cognitive anomalies or overly salient visual artifacts introduced during generation but outside the intended design scope. Edge cases undergo multiple rounds of review to ensure annotation accuracy and dataset consistency. This manual process complements automated checks and preserves the dataset's integrity and intended difficulty. ***(iii) Category balancing.*** We use stratified sampling to balance forgery types and credential formats, preventing evaluation bias due to category skew. ***(iv) Task formatting.*** Each forged sample is paired with a QA instance that asks either for identification of the forged region or of the forgery type. All localization tasks are cast into multiple-choice format to ensure evaluation consistency. Each finalized sample is stored as a structured record comprising: image, credential type, forgery category, forgery region, and reasoning statement.

Ethical Considerations. To fully mitigate any risk of privacy breaches, we have replaced all fields and information from the original samples. All personal information presented in the credentials, including names, addresses, and identification numbers, is entirely fictitious and does not correspond to any real individuals. Therefore, our dataset poses no risk of personal data leakage.

4 CogniCredBench

4.1 Experimental Setup

CogniCredBench evaluates a diverse set of latest open-source MLLMs including the Qwen3-VL series [27], InternVL3.5 series [37], Ovis2.5 series [15], GLM-4.5V [31], MiniCPM-V4_5 [42],Step-VL-10B [9], and DeepSeek-VL2 [32]. Thinking models include Qwen3-VL series and MiniCPM-o-2_6 [42]. We also evaluate three leading proprietary MLLMs: GPT-4o [17], Gemini-3-Pro [6], and Claude-Opus-4 [1].

Our evaluation consists of two complementary tasks: a closed-set classification task and an open-ended reasoning task, both guided by structured prompts which are provided in the appendix. These tasks are designed to assess the ability of a model to identify cognitive-level forgeries in credential images.

Task 1: Closed-set Classification Task. In this task, each evaluated model receives an instruction that enumerates the candidate forgery types and briefly provides the corresponding introduction. Under this instruction, the model is required solely to reason and select from the predefined, mutually exclusive options and present its prediction in the prescribed output format[5]. As the task is closed-form with an explicitly defined target space, we assess performance using strict accuracy: a prediction is deemed correct only if it matches the ground-truth forgery subtype.

Task 2: Open-ended Reasoning Task. To assess whether models can autonomously discover cognitive anomalies, we propose an open-ended reasoning task that evaluates their ability to identify such anomalies through autonomous exploration. In contrast to the closed-set classification setting, models are provided only with the credential image and a concise, abstract instruction. Each model must independently execute the entire reasoning pipeline[6]: (i) determine whether a forgery is present, (ii) identify the nature of the underlying cognitive anomalies, and (iii) produce a coherent explanation within its `<think>` block.

Given the open-ended nature of the responses, string matching alone is inadequate, as models may accurately express the same logical flaw using different terminology. Therefore, we adopt an LLM-as-a-Judge evaluation paradigm (using GPT-4o), which examines both the model's `<answer>` verdict and its corresponding `<think>` reasoning, comparing them against the ground-truth annotation. A response is deemed correct only if the judge determines that the model has successfully identified the core cognitive anomaly and its reasoning aligns with the annotated reasoning statement.

4.2 Main Results

Table 1 presents the consolidated results of CogniCredBench, revealing substantial challenges for current MLLMs and highlighting clear specialization across models. Particularly, we have the following key observations:

[5] `<think>...</think><type>...</type>`

[6] `<think>...</think><answer>...</answer>`

Model	Internal Reasoning		World Knowledge		Visual Reasoning		Contextual Reasoning		Format Compliance	
	Obj. (T1)	Subj. (T2)	Obj. (T1)	Subj. (T2)	Obj. (T1)	Subj. (T2)	Obj. (T1)	Subj. (T2)	Obj. (T1)	Subj. (T2)
Closed-source Models										
GPT-4o	60.60	41.99	45.38	53.69	20.05	1.25	52.80	23.75	19.38	4.34
Claude-opus-4	86.93	80.52	60.00	62.85	42.54	10.72	89.00	78.80	44.16	18.49
Gemini-3-Pro	**97.81**	89.78	83.50	83.14	**65.09**	**44.54**	98.50	**95.40**	**56.20**	**71.98**
Open-source Models										
GLM-4.5V	93.60	67.78	70.71	51.00	32.61	7.74	**99.20**	30.10	29.48	5.56
InternVL3_5-8B	72.41	41.15	35.40	48.72	34.00	12.43	73.35	37.05	30.44	16.95
InternVL3_5-14B	75.56	66.49	82.61	58.04	32.85	10.66	72.75	42.35	40.41	22.94
InternVL3_5-38B	82.63	67.27	63.10	50.31	34.45	7.58	63.30	45.85	44.39	13.20
InternVL3_5-30B-A3B	72.59	36.61	70.63	14.91	30.75	3.79	60.60	33.15	24.05	4.90
InternVL3_5-241B-A28B	82.07	70.00	62.41	53.20	34.97	8.57	91.20	53.75	51.37	19.64
Qwen3-VL-235B-A22B-Instruct	16.50	80.34	59.71	53.60	14.73	16.77	41.35	59.40	21.78	26.92
Qwen3-VL-30B-A3B-Instruct	74.39	69.07	34.66	52.91	26.56	4.82	64.45	34.00	43.24	13.29
Qwen3-VL-32B-Instruct	85.03	**90.74**	53.89	**84.32**	39.54	29.46	56.85	84.55	45.21	36.53
Qwen3-VL-8B-Instruct	73.82	62.04	**84.36**	56.90	25.11	10.46	74.45	43.75	36.00	19.78
Ovis2.5-2B	52.30	41.06	54.95	37.35	38.41	10.52	59.75	15.45	39.19	2.63
Ovis2.5-9B	70.81	74.81	63.58	52.14	28.42	12.13	63.65	38.55	35.87	12.11
DeepSeek-VL2	53.95	19.51	44.11	29.33	30.69	6.81	38.95	3.45	19.91	5.89
MiniCPM-V4_5	77.37	77.01	57.56	45.46	32.53	13.78	64.35	68.05	32.44	10.07
Step3-VL-10B	50.77	68.44	30.39	57.72	8.91	7.19	13.35	26.95	17.34	26.62
Qwen3-VL-32B-Thinking	81.30	76.77	38.94	81.38	26.99	8.79	85.80	87.90	32.84	42.12
Qwen3-VL-30B-A3B-Thinking	73.97	70.67	31.12	60.03	20.03	1.97	75.25	66.75	40.80	31.72
MiniCPM-o-2_6	56.84	20.26	29.37	19.59	13.04	1.13	17.85	0.30	15.51	2.44
Human Score	98.12	–	97.09	–	98.45	–	99.51	–	97.08	–

Table 1: Consolidated Model Accuracy (ACC %) on the Five Reasoning Capabilities, comparing Closed-set Classification Task (Objective, T1) and Open-ended Reasoning Task (Subjective, T2). The highest score in each column is highlighted in bold. "–" is because it is difficult for humans to perform subjective evaluations.

① *Most evaluated MLLMs exhibit a substantial performance drop on the open-ended reasoning task compared with the closed-set classification task.* In many categories, models that achieve strong T1 accuracy collapse to less than half of that performance in T2, and in the hardest capabilities—particularly Visual Reasoning and Format Compliance—the gap can exceed 40 percentage points. This indicates that current MLLMs are more robust when operating under clear evaluation criteria and a constrained output space, yet still struggle with the semantic openness and autonomous reasoning required in T2. Nevertheless, the strong correlation between performance trends on T1 and T2 suggests that the closed-set task design serves as a reliable proxy for assessing a model's overall reasoning capability.

② *Persistent Human–Model Disparity* Human annotators achieve near-perfect accuracy across all five reasoning dimensions, indicating that the benchmark tasks are clear, solvable, and free of ambiguity. In contrast, all evaluated models perform substantially worse than humans—most notably in Visual Reasoning,

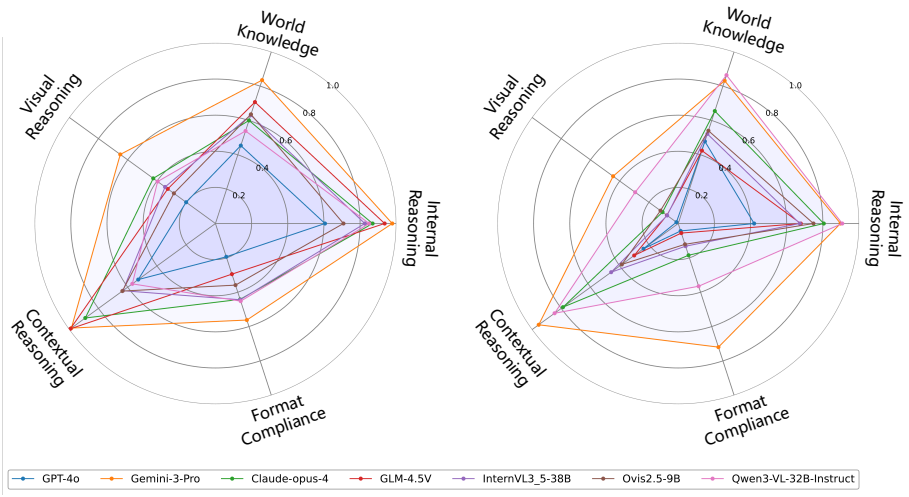

Fig. 4: Radar chart of T1(left)&T2(right) task performance for selected large models across five reasoning capabilities. The results show that while models generally perform well on Internal Reasoning and World Knowledge, their abilities in Visual Reasoning and Format Compliance remain markedly weaker.

where human accuracy reaches 98.45%, while the best models score below 45% on T1 and often fall under 30% on T2. This pronounced human–model gap indicates that current systems still have substantial room for improvement in detecting and reasoning about cognitive-level forgeries—not only in identifying semantic inconsistencies but especially in scenarios that require visual cues. Logical anomalies grounded in visual structure often involve spatial relations, layout constraints, and high-level visual paradigms, which remain fundamental weaknesses of today's multimodal models. This highlights a large and still-unresolved space for advancing visual–structural anomaly detection, pointing to a core bottleneck in the development of robust multimodal cognitive reasoning.

③ *Among the five capabilities, Visual Reasoning and Format Compliance emerge as consistent weaknesses across nearly all models.* As shown in Figure 4, even strong performers in other dimensions struggle considerably in these two areas. For example, while GLM-4.5V reaches 99.20% accuracy on closed-set Contextual Reasoning, its Visual Reasoning and Format Compliance scores drop to 32.61% and 29.48% in T1, and further decline to 7.74% and 5.56% in T2. Similar patterns are observed across most models: many achieve 60–90% accuracy on Internal Reasoning or Contextual Reasoning, yet fall below 30%—and frequently into single digits—in Visual Reasoning and Format Compliance under open-ended evaluation. This pattern suggests that when confronted with text-rich images, models tend to allocate disproportionate attention to textual content, resulting in impairing their visual logic reasoning.

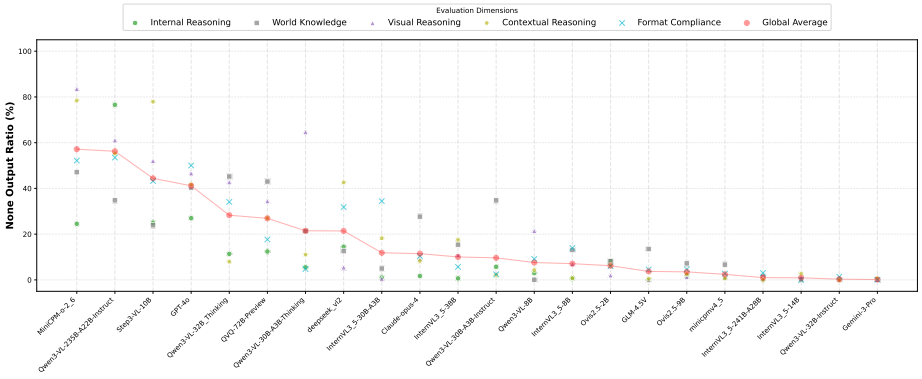

Fig. 5: The figure displays the proportion of samples predicted as "non-forged" or "anomaly-free" by the models.

④ *Current text-centric reasoning paradigms face substantial limitations in deep visual cognition*, Table 1 demonstrate that using the Qwen3-VL series as a representative case, while the "Thinking" mechanism effectively bolsters Contextual Reasoning—with the 32B model surging from 56.85% to 85.80% and the 30B-A3B model improving from 64.45% to 75.25%—it paradoxically degrades Visual Reasoning performance. Specifically, accuracy in this dimension drops from 39.54% to 26.99% for the 32B version, and from 26.56% to 20.03% for the 30B-A3B version. This performance divergence reveals a fundamental mechanistic flaw in the inference process: rather than conducting genuine cognitive steps at the image level, these models merely extract preliminary visual features and subject them to isolated, repetitive text-based deduction. Consequently, true multimodal visual reasoning does not occur, indicating a severe decoupling of internal textual logic from visual grounding.

⑤ *Refusal-Driven Performance Degradation* When confronted with challenging tasks—particularly Visual Reasoning—models often adopt a conservative evasion strategy, defaulting to "anomaly-free" outputs to avoid uncertain generation. This phenomenon directly accounts for the anomalous underperformance of models like GPT-4o and MiniCPM-o-2_6 in our primary evaluation Table 1, as they exhibit exceptionally high refusal rates Figure 5. Notably, Visual Reasoning incurs one of the highest refusal rates across all evaluated dimensions, a tendency that potentially masks the models' genuine reasoning capabilities.

4.3 Analysis of Counter Intuitive Results

GPT-4o's Failure on Visual and Structural Forgery Detection. GPT-4o exhibits strikingly low performance on the T2 open-ended reasoning task for Visual Reasoning (1.25%) and Format Compliance (4.34%). In the representative visual forgery cases shown in the blue-shaded region of Figure 6—such as two different ID photos appearing on the same passport, abnormally large gaps be-

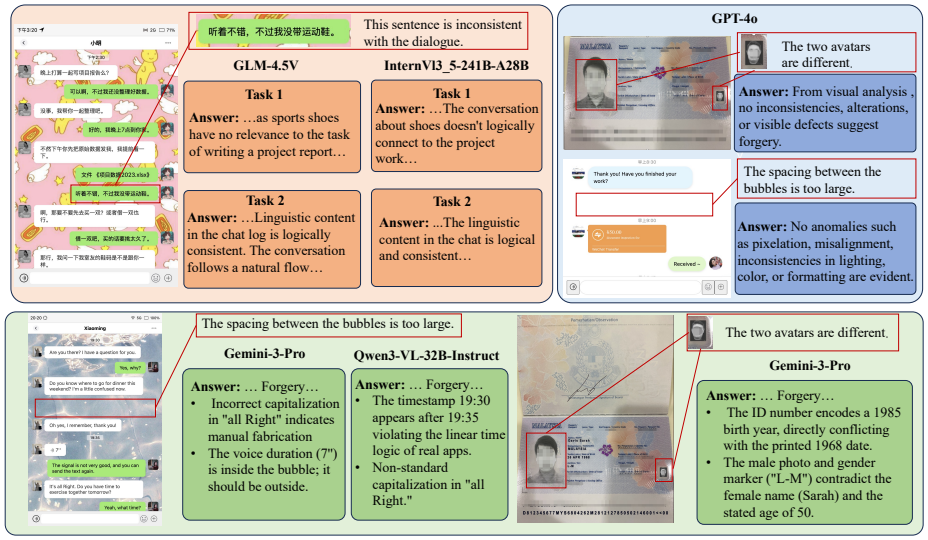

Fig. 6: Analysis of Counter-Intuitive Results. Representative failure cases categorized by background color: (Orange) Performance collapse from T1 to T2 in Contextual Reasoning; (Blue) GPT-4o's visual perception bias, overlooking macro-level structural inconsistencies (e.g., mismatched photos, abnormal spacing); (Green) "Textual overriding," where models ignore glaring visual anomalies to over-analyze benign text features. Red boxes highlight actual inconsistencies.

tween chat bubbles—the model repeatedly outputs `[No Forgery]`, even though its `<think>` trace indicates that it examined the relevant visual elements. A possible explanation is that GPT-4o's visual processing is heavily skewed toward OCR-style text extraction and local artifact detection, while paying insufficient attention to visual paradigms, layout conventions, and structural consistency. This tendency aligns with observations from recent evaluations of MLLMs [28], which similarly report weaknesses in cross-region visual association, structural reasoning, and layout consistency checking. Such a bias may account for GPT-4o's significant performance collapse on open-ended visual cognitive-level forgery detection tasks.

Visual Neglect Driven by Textual Overriding. As quantitative results show, models struggle severely with Visual Reasoning (e.g., T2 accuracy is merely 44.54 % for Gemini-3-Pro and 29.46 % for Qwen3-VL-32B-Instruct). Qualitative analysis reveals the core bottleneck: a "textual overriding" effect where models prioritize semantic extraction over spatial and structural integrity. As illustrated in the green-shaded region of Figure 6, leading models like Gemini-3-Pro and Qwen3-VL-32B-Instruct completely overlook glaring macro-level visual contradictions, such as abnormal chat-bubble gaps or mismatched portrait photos on an ID document. Instead, they are "hijacked" by textual cues, overanalyzing benign details to report nonexistent issues—such as fabricated cap-

italization mistakes, linear time violations, or forced OCR-derived information conflicts. This demonstrates that current MLLMs lack a genuine perception of layout conventions and cross-region visual consistency, completely deviating from true visual auditing.

Performance Collapse from T1 to T2 in Contextual Reasoning. In the Contextual Reasoning evaluation shown in the orange-shaded region of Figure 6, we conduct a qualitative analysis using the illustrated dialogue. The conversation is semantically coherent: the speakers discuss project data, file transfer, and meeting coordination before casually mentioning not bringing sports shoes. When framed as an objective multiple-choice question, both GLM-4.5V and InternVL3.5-241B-A28B achieve near-perfect accuracy in identifying this sentence as inconsistent with the project-related context, showing strong performance in constrained semantic discrimination. However, when the same scenario is reformulated as an open-ended reasoning task requiring the model to independently detect the inconsistency and articulate why it is incoherent, the models exhibit a sharp performance collapse. Typical failures include offering only superficial descriptions, omitting the explicit semantic break, or even rationalizing the dialogue as coherent. This case reveals the core bottleneck: such tasks demand modeling of pragmatic context and constructing cross-sentence causal chains—capabilities that extend far beyond simple semantic matching. Furthermore, this discrepancy suggests that while MLLMs can recognize statistical anomalies within a provided candidate space, they struggle to autonomously construct a global cognitive map of the conversation. Without the structural scaffolding of predefined options, the models' internal reasoning often drifts toward local textual patterns, failing to maintain the high-level situational awareness necessary for robust cognitive-level forgery detection.

5 Discussion and Conclusion

In this paper, we introduce CogniCred, the first cognitive-focused dataset, which includes over 16,000 forged credential samples with 5 major categories and 14 subcategories. Based on this dataset, we further build CogniCredBench to evaluate models' abilities to detect cognitive-level credential forgeries across two distinct tasks. Experimental results reveal pronounced performance disparities among representative MLLMs, confirming the benchmark's discriminative power. Notably, our analysis shows that many models rely heavily on textual reasoning while overlooking salient visual inconsistencies, exposing a critical gap in current multimodal understanding and vision grounding. To facilitate future research, the CogniCred dataset, benchmark, and evaluation code will be made publicly available. We hope our work will draw the community's attention to complex credential forgeries and further advance academic research in the security field.

Acknowledgments This work is supported by the National Natural Science Foundation of China (No. 62372408). This work was also supported by Ant Group.

References

1. Anthropic: Claude Opus 4.1. https://www.anthropic.com/claude/opus (2025), accessed: 2025-08-05
2. Bertrand, R., Terrades, O.R., Gomez-Krämer, P., Franco, P., Ogier, J.M.: A conditional random field model for font forgery detection. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 576–580. IEEE (2015)
3. Chen, C., McCloskey, S., Yu, J.: Image splicing detection via camera response function analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5087–5096 (2017)
4. Chen, Z., Chen, S., Yao, T., Sun, K., Ding, S., Lin, X., Cao, L., Ji, R.: Enhancing tampered text detection through frequency feature fusion and decomposition. In: European Conference on Computer Vision. pp. 200–217. Springer (2024)
5. Dong, L., Liang, W., Wang, R.: Robust text image tampering localization via forgery traces enhancement and multiscale attention. IEEE Transactions on Consumer Electronics (2024)
6. Google: Gemini 3 pro api. https://deepmind.google/technologies/gemini/ (2026), accessed: 2026-03
7. Guan, H., Wang, Y., Xie, L., Nag, S., Goel, R., Swamy, N.E.N., Yang, Y., Xiao, C., Prisby, J., Maciejewski, R., et al.: Idnet: A novel dataset for identity document analysis and fraud detection. arXiv preprint arXiv:2408.01690 (2024)
8. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020). https://doi.org/10.1109/TIP.2020.2967770
9. Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J., Hu, J., Lin, K., Zhao, L., Huang, M., Yuan, S., Qu, W., Wang, X., Lai, Y., Zhao, Y., Zhang, Y., Shi, Y., Chen, Y., Weng, Z., Meng, Z., Li, A., Kong, A., Dong, B., Wan, C., Wang, D., Qi, D., Li, D., Yu, E., Li, G., Yin, H., Zhou, H., Zhang, H., Yan, H., Zhou, H., Peng, H., Zhang, J., Lv, J., Fu, J., Cheng, J., Zhou, J., Yin, J., Xie, J., Wu, J., Zhang, J., Liu, J., Tan, K., Yan, K., Chen, L., Chen, L., Li, M., Zhao, Q., Sun, Q., Pang, S., Fan, S., Shang, S., Zhang, S., You, T., Ji, W., Xie, W., Yang, X., Hou, X., Jiao, X., Ren, X., Kong, X., Huang, X., Wu, X., Chen, X., Wang, X., Zhang, X., Wei, Y., Li, Y., Xu, Y., Shen, Y., Peng, Y., Peng, Y., Zhou, Y., Li, Y., Yang, Y., Zhang, Y., Xie, Z., Huang, Z., Lu, Z., Fan, Z., Cheng, Z., Jiang, D., Han, Q., Zhang, X., Zhu, Y., Ge, Z.: Step3-vl-10b technical report (2026), https://arxiv.org/abs/2601.09668
10. Huang, Z., Xia, B., Lin, Z., Mou, Z., Yang, W., Jia, J.: Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant (2024), https://arxiv.org/abs/2408.10072
11. Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C.: Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision **130**(8), 1875–1895 (2022)

12. Laban, P., Schnabel, T., Bennett, P.N., Hearst, M.A.: Summac: Re-visiting nli-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics **10**, 163–177 (2022)
13. Li, Y., Liu, X., Wang, X., Lee, B.S., Wang, S., Rocha, A., Lin, W.: Fakebench: Probing explainable fake image detection via large multimodal models (2024), https://arxiv.org/abs/2404.13306
14. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10770–10780 (2024)
15. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., Han, Y., Li, H., Chen, W., Tang, J., Hou, C., Du, Z., Zhou, T., Zhang, W., Ding, H., Li, J., Li, W., Hu, G., Gu, Y., Yang, S., Wang, J., Sun, H., Wang, Y., Sun, H., Huang, J., He, Y., Shi, S., Zhang, W., Zheng, G., Jiang, J., Gao, S., Wu, Y.F., Chen, S., Chen, Y., Chen, Q.G., Xu, Z., Luo, W., Zhang, K.: Ovis2.5 technical report (2025), https://arxiv.org/abs/2508.11737
16. Luo, D., Liu, Y., Yang, R., Liu, X., Zeng, J., Zhou, Y., Bai, X.: Toward real text manipulation detection: New dataset and new solution. Pattern Recognition **157**, 110828 (2025)
17. OpenAI: Gpt-4o (2024), https://openai.com/index/hello-gpt-4o/, accessed: 1, 6, 7
18. Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., Tao, D.: Deepfake generation and detection: A benchmark and survey (2024), https://arxiv.org/abs/2403.17881
19. Qu, C., Jin, L., Li, J., Liu, J., Yu, B., Xie, J., Liu, J.: Dino-mac: First-place winner solution of the cvpr2026 robust deepfake detection challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2406–2415 (2026)
20. Qu, C., Liu, C., Liu, Y., Chen, X., Peng, D., Guo, F., Jin, L.: Towards robust tampered text detection in document image: New dataset and new solution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5937–5946 (2023)
21. Qu, C., Liu, J., Chen, H., Yu, B., Liu, J., Wang, W., Jin, L.: Explainable tampered text detection via multimodal large models. arXiv preprint arXiv:2412.14816 (2024)
22. Qu, C., Zhong, Y., Guo, F., Jin, L.: Revisiting tampered scene text detection in the era of generative ai. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 694–702 (2025)
23. Qu, C., Zhong, Y., Guo, F., Jin, L.: Omni-iml: Towards unified interpretable image manipulation localization. In: The Fourteenth International Conference on Learning Representations (2026)
24. Qu, C., Zhong, Y., Liu, C., Xu, G., Peng, D., Guo, F., Jin, L.: Towards modern image manipulation localization: A large-scale dataset and novel methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781–10790 (2024)
25. Qu, C., Zhong, Y., Liu, J., Zhu, X., Yu, B., Jin, L.: Textshield-r1: Reinforced reasoning for tampered text detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 8621–8629 (2026)
26. Qu, C., Zhong, Y., Zhu, X., Li, J., Jiang, C., Jin, L., et al.: Detect any ai-counterfeited text image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 35437–35450 (2026)

27. Qwen Team: Qwen3-VL: The Most Powerful Vision-Language Model in the Qwen Series. https://github.com/QwenLM/Qwen3-VL (2025), accessed: 2025-11-13
28. Ramachandran, R., Garjani, A., Bachmann, R., Atanov, A., Kar, O.F., Zamir, A.: How well does gpt-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks (2025), https://arxiv.org/abs/2507.01955
29. Shao, H., Huang, K., Wang, W., Huang, X., Wang, Q.: Progressive supervision for tampering localization in document images. In: International Conference on Neural Information Processing. pp. 140–151. Springer (2023)
30. Shao, H., Qian, Z., Huang, K., Wang, W., Huang, X., Wang, Q.: Delving into adversarial robustness on document tampering localization. In: European Conference on Computer Vision. pp. 290–306. Springer (2024)
31. Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., Tang, J.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), https://arxiv.org/abs/2507.01006
32. *et al.*, Z.W.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding (2024), https://arxiv.org/abs/2412.10302
33. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: Fever: a large-scale dataset for fact extraction and verification. arXiv preprint arXiv:1803.05355 (2018)
34. Van Beusekom, J., Shafait, F., Breuel, T.M.: Text-line examination for document forgery detection. International Journal on Document Analysis and Recognition (IJDAR) **16**, 189–207 (2013)
35. Wang, J., Lv, C., Li, X., Dong, S., Li, H., Li, C., Shao, W., Luo, P., et al.: Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models. arXiv preprint arXiv:2503.15024 (2025)
36. Wang, S., Lin, H., Luo, Z., Ye, Z., Chen, G., Ma, J.: Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. arXiv preprint arXiv:2406.11288 (2024)
37. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265
38. Wang, Y., Xie, H., Xing, M., Wang, J., Zhu, S., Zhang, Y.: Detecting tampered scene text in the wild. In: European Conference on Computer Vision. pp. 215–232. Springer (2022)

39. Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., Li, W.: Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv preprint arXiv:2503.14905 (2025)
40. Xie, Y., Zhang, J., Chen, P., Wang, Z., Wang, W., Gao, L., Li, P., Sun, H., Zhang, Q., Qiao, Q., Fan, J., Lian, Z.: Textflux: An ocr-free dit model for high-fidelity multilingual scene text synthesis (2025), https://arxiv.org/abs/2505.17778
41. Yan, Q., Fan, Y., Li, H., Jiang, S., Zhao, Y., Guan, X., Kuo, C.C., Wang, X.E.: Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. arXiv preprint arXiv:2502.16033 (2025)
42. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
43. Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., Wu, Z., Chen, Y., Lin, D., He, C., Li, W.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models (2025), https://arxiv.org/abs/2410.09732
44. Zhang, Y., Colman, B., Guo, X., Shahriyari, A., Bharaj, G.: Common sense reasoning for deepfake detection (2024), https://arxiv.org/abs/2402.00126
45. Zhu, F., Liu, Z., Ng, X.Y., Wu, H., Wang, W., Feng, F., Wang, C., Luan, H., Chua, T.S.: Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. arXiv preprint arXiv:2410.21311 (2024)