Real5-OmniDocBench: A Full-Scale Physical Reconstruction Benchmark for Robust Document Parsing in the Wild

Cheng Cui[1⋆], Changda Zhou[1⋆], Ziyue Gao[1,2], Xueqing Wang[1], Tingquan Gao[1], Jing Tang[2], and Yi Liu[1⋆⋆]

[1] PaddlePaddle Team, Baidu Inc., Beijing, China paddleocr@baidu.com
[2] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

Abstract. While Vision-Language Models (VLMs) achieve near-perfect scores on digital document benchmarks like OmniDocBench, their performance in the unpredictable physical world remains largely unknown due to the lack of controlled yet realistic evaluations. We introduce Real5-OmniDocBench, the first benchmark to provide a full-scale, one-to-one physical reconstruction of the complete OmniDocBench v1.5 test set (1,355 images) across five critical real-world scenarios: Scanning, Warping, Screen-Photography, Illumination, and Skew. Unlike prior benchmarks that either lack digital correspondence or employ partial sampling, our complete ground-truth mapping enables, for the first time, rigorous factor-wise attribution of performance degradation, allowing us to pinpoint whether failures stem from geometric distortions, optical artifacts, or model limitations. Our benchmark establishes a challenging new standard for the community, demonstrating that the "reality gap" in document parsing is far from closed, and provides a diagnostic tool to guide the development of truly resilient document intelligence.
The Real5-OmniDocBench dataset is publicly available at `https://huggingface.co/datasets/PaddlePaddle/Real5-OmniDocBench`.

Keywords: Document Parsing · Real-world Benchmark · Vision-Language Models

1 Introduction

Document parsing, defined as the task of transforming unstructured document images into structured formats such as Markdown and LaTeX, has become a critical benchmark for evaluating the fine-grained visual reasoning capabilities of Vision-Language Models (VLMs). Recent progress has been largely driven by high-quality benchmarks like OmniDocBench [18], which provide precise annotations across diverse document types and enable standardized evaluation. However, these benchmarks predominantly rely on "born-digital" documents captured under ideal, distortion-free conditions. The uncomfortable truth is that a

⋆ Equal contribution.
⋆⋆ Corresponding author.

model scoring extremely high on OmniDocBench today might fail catastrophically when faced with real-world documents, such as a page curved by a book spine, a receipt photographed under a desk lamp, or a screenshot corrupted by moiré patterns.

In real-world deployment, document images inevitably suffer from complex physical perturbations: non-rigid warping from binding, perspective distortions from handheld capture, non-uniform illumination from ambient light, and optical artifacts from secondary screen-photography. Despite anecdotal evidence of such vulnerabilities, the community lacks a systematic understanding of how and why models fail under these conditions. Prior evaluation efforts fall into two categories, both insufficient: uncontrolled in-the-wild datasets (*e.g.*, WildDoc [22]) capture realistic degradations but lack digital ground-truth correspondence, making it impossible to precisely measure the impact of real-world environmental interference on model reasoning; partial physical simulations (*e.g.*, DocPTBench [7]) offer controllability but employ coarse-grained sampling that fails to capture the full spectrum of real-world distortions or enable rigorous factor-wise diagnosis. What is missing is a benchmark that combines the realism of physical capture with the precision of controlled one-to-one digital mapping—a tool that can tell us not just that a model failed but also exactly which distortion caused the failure.

We introduce Real5-OmniDocBench to fill this critical gap. Our core innovation is simple yet powerful: we perform a full-scale, one-to-one physical reconstruction of the entire OmniDocBench v1.5 test set, consisting of 1,355 pages, across five meticulously designed physical scenarios, including Scanning, Warping, Screen-Photography, Illumination, and Skew. For each digital source image, we produce five physical variants using professional-grade printing and heterogeneous mobile capture devices while inheriting the complete ground-truth annotations from the original. This design transforms physical distortions from uncontrolled confounders into controlled, independent variables. Consequently, our benchmark enables controlled factor-wise analysis of the domain gap between ideal conditions and real-world physical scenarios, allowing precise evaluation of performance variations across different scenarios under fully comparable settings. This achieves a level of comparability that was previously unattainable.

Our contributions are threefold:

- **First Full-Scale Physical Benchmark for Controlled Robustness Analysis.** We present Real5-OmniDocBench, the first benchmark to offer a complete 1,355-image physical reconstruction with strict one-to-one digital correspondence. Unlike partial or uncontrolled datasets, our design enables rigorous attribution of performance degradation to specific physical factors, providing the community with a diagnostic tool rather than just another leaderboard.
- **Comprehensive Multi-Scenario Physical Modeling.** We systematically construct five orthogonal physical scenarios: Scanning, Warping, Screen-Photography, Illumination, and Skew, each built from multiple diverse sub-conditions (*e.g.*, Warping: folding, cylinder, crumpling, corner, book). This

scenario-level design enables precise diagnosis of model robustness—for instance, revealing that a model resilient to scanning artifacts degrades severely under handheld warping.
– **Large-Scale Empirical Study Revealing Counter-Intuitive Insights.** Through comprehensive benchmarking of 17 state-of-the-art models, we uncover a finding that challenges the prevailing scaling paradigm: compact, domain-specialized VLMs (*e.g.*, PaddleOCR-VL-1.6 at 0.9B parameters) consistently outperform much larger generalist models under physical stress. This suggests that robustness to real-world distortions is not a simple function of parameter count, but requires domain-specific inductive biases—a crucial insight for future model design.

By publicly releasing Real5-OmniDocBench, we provide the community with a much-needed stress test for document intelligence. Our results establish a demanding baseline that reveals the substantial gap between digital perfection and real-world reliability. We hope this benchmark will catalyze research toward models that truly understand documents in both pristine archives and the messy, unpredictable physical world.

2 Related Work

Evolution of Document Parsing Benchmarks. Document understanding benchmarks have evolved from early single-task OCR evaluations, such as the ICDAR series [10] focusing on text detection and recognition, to complex end-to-end structured parsing. In this progression, OmniDocBench [18] has established itself as a core benchmark for evaluating the document parsing capabilities of VLMs (*e.g.*, Qwen3-VL [1], Gemini 3 Pro [9]), owing to its definition of nine document types and a three-tier evaluation framework (full-page, module, and attribute levels). Although OmniDocBench provides high-precision annotations, it predominantly relies on "born-digital" images, overlooking the physical imaging degradations encountered during real-world document acquisition.

Current Landscape of VLM Evaluation Sets. To further explore the parsing limits of VLMs, several specialized benchmarks have recently emerged. OlmOCR-Bench [20] introduces unit-test-driven verification for the structural equivalence of formulas and tables, while OceanOCR-Bench [3] focuses on text-image association and cross-modal alignment in complex multimodal layouts. For real-world document understanding, WildDoc [22] collects documents captured in natural environments, challenging models with diverse lighting conditions and complex backgrounds. However, because these images are not paired with corresponding pristine digital originals, they are less suited for controlled comparisons between ideal and real-world acquisition conditions. Concurrent efforts have also explored physical document acquisition and realistic degradation modeling. DocPTBench [7] investigates photographed document parsing through real image collection, whereas another concurrent work [11] studies document parsing using realistic scene synthesis and document-aware training. These efforts

provide valuable complementary perspectives for evaluating document parsing under realistic conditions. Real5-OmniDocBench complements these benchmarks by focusing on complete digital–physical correspondence under controlled acquisition settings.

Uniqueness of Real5-OmniDocBench. Distinct from contemporary efforts based on synthetic rendering [11] or subset-based physical acquisition [7], our work achieves one-to-one physical reconstruction of the entire OmniDocBench corpus (1,355 source digital images), yielding 6,775 fully aligned digital–physical pairs under controlled real-world conditions. This rigorous paired design enables controlled factor-wise attribution of performance degradation across five physical factors, namely Scanning, Warping, Screen-Photography, Illumination, and Skew, facilitating more precise diagnosis of model robustness under real-world document acquisition. Together, these characteristics provide a complementary capability for systematically analyzing robustness under controlled physical degradations.

3 The Real5-OmniDocBench Benchmark

3.1 Design Principles and Overall Architecture

Real5-OmniDocBench is constructed based on OmniDocBench v1.5 [18], which contains 1,355 pages across nine document types (academic papers, books, notes, financial reports, magazines, etc.). We follow the one-to-one mapping principle: each original sample corresponds to five real-world scenario variants, totaling 6,775 test samples. The GT inheritance strategy ensures zero-modification reuse of OmniDocBench's JSON annotations (layout, tables, formulas, text, reading order), which is the core design guaranteeing strict comparability of evaluation. Figure 1 illustrates representative examples from our dataset, showcasing the diversity of document types and real-world degradations covered.

3.2 Physical Data Acquisition and Standardization

To ensure the scientific integrity and reproducibility of Real5-OmniDocBench, we implement a rigorous physical data acquisition pipeline. All physical samples are reproduced using a professional-grade Canon C5840 multi-functional printer, configured at 1200 dpi for both color printing and scanning. This high-resolution reproduction ensures that any subsequent parsing degradation is strictly attributable to environmental perturbations rather than source quality loss.

To maintain the structural fidelity of the original digital corpus, we adopt a scale-aware printing strategy: documents are categorized into A3 and A4 formats based on their native digital dimensions before printing. This approach preserves the original font-size ratios and layout densities, which are crucial for precise structural parsing evaluation. For the four handheld capture scenarios (*i.e.*, Warping, Screen-Photography, Illumination, and Skew), we utilize a heterogeneous hardware matrix comprising mainstream mobile devices from Apple,

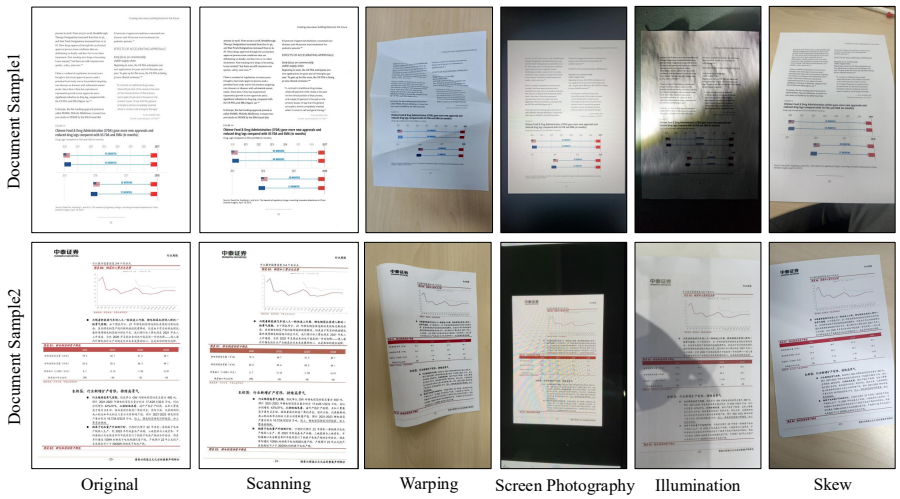


Fig. 1: Representative examples from Real5-OmniDocBench.

Xiaomi, and OPPO. This diverse device selection is instrumental in capturing a wide spectrum of ISP (Image Signal Processing) characteristics, sensor noise profiles, and lens distortions, thereby reflecting the variability encountered in real-world document acquisition across different hardware ecosystems. Apple, Xiaomi, and OPPO devices are used at an approximate ratio of 3:1:1, reflecting a diverse range of consumer hardware during data acquisition.

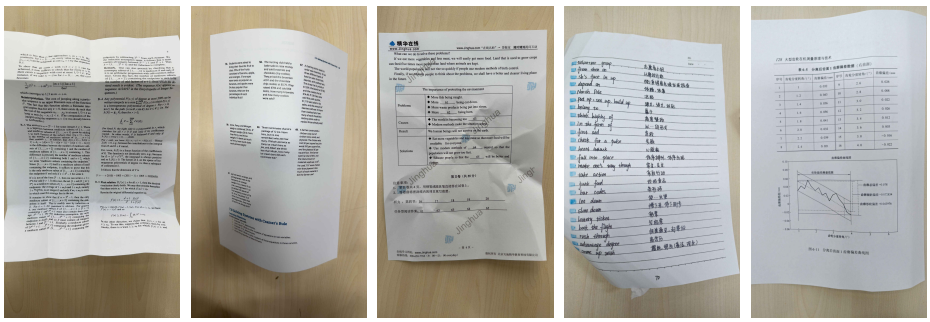
Crucially, our quantitative analysis confirms that this collection strategy yields a substantial and continuous physical distortion spectrum across the corpus. In the Illumination scenario, the mean grayscale value drops to 123.44 ± 26.19, which corresponds to an average reduction of approximately 46% in overall brightness compared to the pristine digital references. For geometric variations, the surface non-rigidity introduced in the Warping scenario achieves an average surface deformation of $4.43\times$ compared to the original flat planes.

Scanning. The Scanning scenario evaluates parsing performance under controlled high-resolution conditions while introducing authentic paper textures and digitization noise. As illustrated in Fig. 2, we implement five representative configurations: *Standard* (a) for ideal, well-aligned capture; *Low-quality* (b) simulating resolution loss from entry-level hardware via iterative print-scan cycles; *Slanted* (c) capturing non-orthogonal alignment common in manual flatbed feeding; *Stapled* (d) featuring local shadows and occlusions from corner clips; and *Bound* (e) simulating book-style scanning with characteristic edge curvature. These settings offer a spectrum of digitization artifacts found in both archival and office environments.

(a) Standard (b) Low-quality (c) Slanted (d) Stapled (e) Bound

Fig. 2: Representative examples of the Scanning scenario.

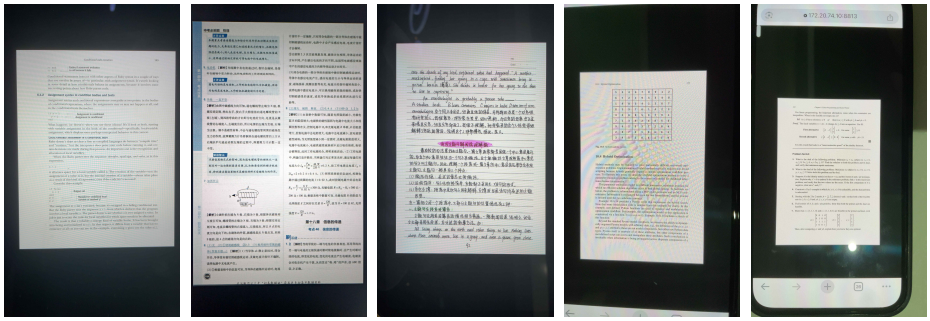
Warping. The Warping scenario simulates non-rigid physical deformations encountered during document handling. We define five typical deformation patterns in Fig. 3: *Folding* (a) creating creases along the centerline; *Cylinder* (b) producing residual curvature from rolling; *Crumpling* (c) introducing dense, stochastic local folds; *Corner* (d) simulating peripheral dog-ears or page curling; and *Book* (e) replicating the natural arc at a binding spine. This diverse set of geometric distortions provides a rigorous test for the model's spatial robustness.



(a) Folding (b) Cylinder (c) Crumpling (d) Corner (e) Book

Fig. 3: Representative examples of the Warping scenario.

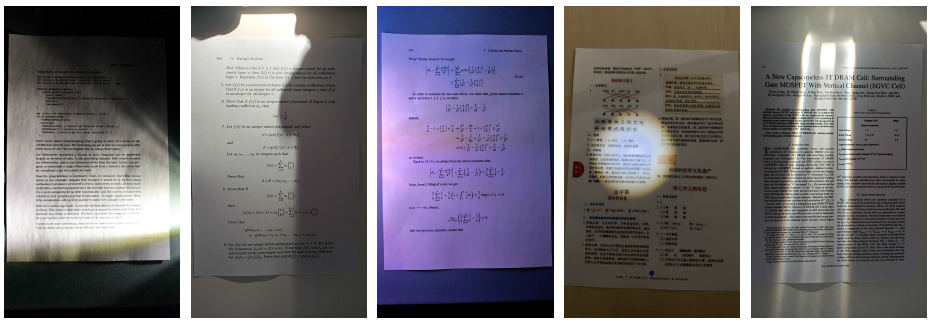
Screen-Photography. This scenario replicates the secondary capture of documents displayed on digital terminals. We categorize the display devices into five distinct terminals as shown in Fig. 4: *Office Monitor* (a) representing standard LCD displays; *Professional Display* (b) featuring high pixel density; *Laptop* (c) utilizing integrated high-resolution panels; *Tablet* (d) simulating portable touchscreen devices; and *Mobile* (e) representing small-scale, high-brightness OLED screens. This selection ensures a systematic assessment across varying pixel structures and moiré patterns.



(a) Office (b) Pro (c) Laptop (d) Tablet (e) Mobile

Fig. 4: Representative examples of the Screen-Photography scenario.

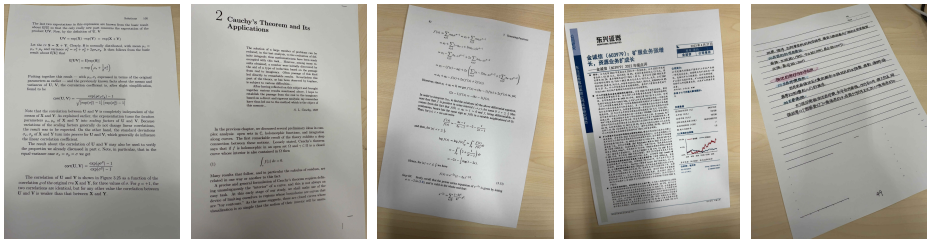
Illumination. The Illumination scenario examines parsing resilience under non-uniform lighting. We construct five complex visual environments in Fig. 5: *Low-light* (a) simulating insufficient luminance; *Shadow* (b) creating partial occlusion; *Color-cast* (c) introducing chromatic shifts; *Flashlight* (d) generating concentrated overexposure; and *Refraction* (e) simulating visual distortions through transparent media. This design probes the model's ability to maintain semantic consistency under significant contrast perturbations.



(a) Low-light (b) Shadow (c) Color-cast (d) Flashlight (e) Refraction

Fig. 5: Representative examples of the Illumination scenario.

Skew. The Skew scenario models 3D perspective distortions inherent in handheld photography. We design five pose configurations in Fig. 6: *Pitch* (a) causing vertical foreshortening; *Roll* (b) producing horizontal trapezoidal distortion; *Yaw* (c) resulting in planar rotation; *Compound* (d) combining multi-axis rotations; and *Extreme* (e) featuring aggressive tilt angles with severe edge distortions. These variations reflect the geometric challenges of real-world mobile document capture.



(a) Pitch (b) Roll (c) Yaw (d) Compound (e) Extreme

Fig. 6: Representative examples of the Skew scenario.

3.3 Quality Audit and Refinement

To guarantee the reliability of Real5-OmniDocBench, we implement a multi-stage quality control pipeline combining automated anomaly detection with expert manual verification.

Machine-Assisted Anomaly Detection. We employ a committee of state-of-the-art (SOTA) VLMs to perform initial inference across all captured images. Samples exhibiting catastrophic performance drops (*e.g.*, zero-score outliers) across the entire committee are flagged as potential anomalies. This phase efficiently locates samples suffering from severe hardware-induced failures, such as extreme motion blur, hardware-level overexposure, or accidental frame misalignment.

Expert Manual Verification and Recapture. All flagged samples undergo a rigorous manual audit. We perform three rounds of full-set inspection to ensure: (1) Semantic Correspondence: Each image correctly maps to its original digital counterpart in OmniDocBench v1.5; (2) Structural Integrity: No significant content is lost due to improper framing or occlusion; (3) Geometric Orthogonality: Image orientation is standardized, correcting any unintentional 90/180-degree rotations. Samples confirmed with irrecoverable artifacts (*e.g.*, unreadable text due to severe defocus) are subjected to a recapture-and-replace loop until they meet our quality threshold.

Fidelity vs. Real-world Degradation. A critical distinction is made between unintentional capture errors and representative real-world degradation. While our 1200 dpi printing ensures source fidelity, handheld capture inherently introduces a resolution gap compared to digital originals. We intentionally retain general clarity degradation, as it reflects the imaging characteristics encountered in real-world mobile document acquisition. Our audit protocol only filters out "unusable" samples (*e.g.*, severe blur causing illegibility), while preserving authentic artifacts like moiré and minor noise. This ensures the benchmark evaluates the model's robustness to real-world deployment conditions rather than its tolerance for accidental photographic errors.

Dataset Completeness. Through continuous refinement, we maintain 100% coverage of the original 1,355 OmniDocBench samples across all five scenarios. The final version of Real5-OmniDocBench strikes a balance between physical realism and evaluation reliability, providing a stable foundation for benchmarking the next generation of VLMs.

Limitations of the Quality Audit. Despite our rigorous audit pipeline, certain subtle edge cases may still persist. Because our anomaly detection primarily filters samples with extremely limited readable information, some partially degraded cases may remain, particularly under challenging Illumination or complex Warping scenarios. For example, a sample might exhibit localized motion blur, minor shadow occlusions, or slight corner clipping while its main body remains legible. Although our manual inspections on these subsets confirm the overall validity and usability of the pipeline, ensuring that every sample remains perfectly readable by human annotators under all extreme real-world conditions remains an ongoing challenge. We recognize fine-grained, per-scenario human-readability evaluation as a limitation of our current pipeline and a valuable direction for future benchmark refinement.

4 Evaluation Methodology and Metrics

Real5-OmniDocBench is designed to be fully compatible with the evaluation ecosystem of OmniDocBench [18], facilitating a rigorous cross-domain diagnostic analysis between pristine digital originals and their corresponding physical degradations.

Core Metrics and Scientific Logic. We adopt the multi-dimensional scoring framework of OmniDocBench, which decouples layout, content, and structural logic. The metrics are defined as follows:

- **Text and Reading Order**: We utilize the Normalized Edit Distance (NED):

$$NED = \frac{Lev(s_1, s_2)}{\max(|s_1|, |s_2|)}$$

  where Lev denotes the Levenshtein distance. This normalized distance is applied to both full-text OCR and reading-order sequences to measure character-level errors and topological inconsistencies, respectively, with lower values indicating better performance.
- **Formula Parsing (CDM)**: A key feature of this framework is the character detection matching (CDM) [24] metric. Unlike string-based metrics, CDM converts formulas into computational graphs to evaluate structural equivalence, ensuring an objective assessment of mathematical logic even when physical noise induces minor syntactic variations.

– **Table Structure and Content (TEDS)**: We employ the Tree Edit Distance-based Similarity (TEDS) [29]:

$$TEDS(G_a, G_p) = 1 - \frac{EditDist(G_a, G_p)}{\max(|G_a|, |G_p|)}$$

where G represents the table's tree structure. We report both **TEDS (Overall)** and TEDS-Struct to identify whether failures occur in content recognition or structural reconstruction.
– **Overall Metric**: The overall performance is quantified as follows:

$$\text{Overall} = \frac{(1 - \text{Text NED}) \times 100 + \text{Table TEDS} + \text{Formula CDM}}{3}$$

Model Array and Alignment Protocol. To ensure the credibility of our benchmark, we categorize test subjects based on their alignment with official records:

– **Officially Aligned Baselines**: We prioritize models verified in the original OmniDocBench report, including Qwen3-VL-235B, Qwen2.5-VL-72B, Gemini-2.5 Pro, MinerU-1.8.2, and PP-StructureV3. Our evaluation principle follows a strict alignment-first protocol: we first perform inference on the official digital test set; only models that successfully reproduce official metrics are then evaluated on our Real5-OmniDocBench dataset. For models where official digital metrics cannot be currently aligned, we withhold their Real5-OmniDocBench scores to maintain evaluation rigor.
– **Supplemental Frontier Evaluations**: We additionally include the latest state-of-the-art models, such as Gemini-3 Pro, GPT-5.2, GLM-OCR, DeepSeek-OCR 2 and PaddleOCR-VL-1.6, using identical configurations to reflect the most recent progress in VLM robustness.For models with newer publicly available versions, we evaluate the latest available checkpoints under the same protocol to ensure up-to-date and fair comparison.

5 Experiments and Analysis

5.1 Main Results Overview

Table 1 presents a comprehensive performance analysis of various models on the Real5-OmniDocBench benchmark. The results indicate that different model architectures exhibit distinct robustness profiles when encountering physical degradations. Notably, the specialized PaddleOCR-VL-1.6 achieves an Overall score of 93.19, demonstrating high competitive efficiency at a 0.9B parameter scale compared to significantly larger counterparts, such as Qwen3-VL-235B (88.90) and Gemini-3 Pro (89.24). Across all five evaluated scenarios, this compact model maintains consistent performance, with metrics ranging from 92.48% in Warping to 94.74% in Scanning. Such stability across diverse physical conditions suggests that domain-specific architectural optimization and exposure to diverse artifacts

Table 1: Comprehensive evaluation on Real5-OmniDocBench. **S**: Scanning, **W**: Warping, **SP**: Screen-Photography, **I**: Illumination, **SK**: Skew. Best results are **bolded** and second-best are underlined.

Model Type	Methods	Params	Overall↑	S↑	W↑	SP↑	I↑	SK↑
Pipeline Tools	Marker-1.8.2 [19]	-	60.10	70.27	58.98	63.65	66.31	41.27
	PP-StructureV3 [6]	-	64.45	84.68	59.34	66.89	73.38	37.98
General VLMs	GPT-5.2 [16]	-	78.66	84.43	76.26	76.75	80.88	75.00
	Qwen2.5-VL-72B [2]	72B	86.92	86.19	87.77	86.48	87.25	86.90
	Gemini-2.5 Pro [8]	-	88.21	89.25	87.63	87.11	87.97	89.07
	Qwen3-VL-235B [1]	235B	88.90	89.43	89.99	89.27	89.27	86.56
	Kimi K2.5 [21]	1T	89.09	89.67	88.86	88.39	89.66	88.86
	Gemini-3 Pro [9]	-	89.24	89.47	88.90	88.86	89.53	89.45
Specialized VLMs	Deepseek-OCR 2 [26]	3B	73.01	89.59	66.53	71.65	76.02	61.28
	Deepseek-OCR [25]	3B	73.99	86.17	67.20	75.31	78.10	63.01
	MinerU2-VLM [17]	0.9B	76.95	83.60	73.73	78.77	80.51	68.16
	MonkeyOCR-pro-3B [13]	3.7B	79.49	86.94	78.90	82.44	84.71	64.47
	Nanonets-OCR-s [14]	3B	84.19	85.52	83.56	84.86	85.01	81.98
	dots.ocr [12]	3B	86.38	86.87	86.01	87.18	87.57	84.27
	PaddleOCR-VL [5]	0.9B	85.54	92.11	85.97	82.54	89.61	77.47
	MinerU2.5 [15]	1.2B	85.61	90.06	83.76	89.41	89.57	75.24
	MinerU2.5-Pro [23]	1.2B	88.94	92.11	88.72	91.29	91.31	81.26
	GLM-OCR [27]	0.9B	90.32	92.67	90.68	91.75	91.12	85.39
	PaddleOCR-VL-1.5 [4]	0.9B	92.05	93.43	91.25	91.76	92.16	91.66
	PaddleOCR-VL-1.6 [28]	0.9B	**93.19**	**94.74**	**92.48**	**92.78**	**93.28**	**92.66**

during training may be more critical for real-world document parsing than raw parameter scaling, which primarily benefits semantic priors rather than geometric and optical invariance.

Within the General VLM category, Gemini-3 Pro shows consistent performance in maintaining semantic coherence. While Qwen3-VL-235B exhibits a specialized advantage in the Warping scenario (89.99%), its performance fluctuates in Screen-Photography (89.27%) and Skew (86.56%) tasks. These fluctuations suggest that increasing parameter scale primarily enhances character recognition accuracy but may not inherently resolve geometric ambiguities introduced by 3D perspective transformations. In contrast, traditional Pipeline Tools show a higher sensitivity to environmental changes. For instance, both PP-StructureV3 and Marker-1.8.2 experience a notable decline in Skew scenarios compared to their Scanning results, primarily because heuristic-based layout analysis often struggles with large-angle perspective distortions.

5.2 Scenario-Level Diagnostic Analysis

To provide a granular quantification of model robustness, we decompose the evaluation into four critical sub-tasks: Full-text OCR (Text^{E}), Formula Recognition ($\text{Formula}^{\text{CDM}}$), Table Reconstruction ($\text{Table}^{\text{TEDS}}$), and Reading Order consistency (RO^{E}). This multi-dimensional analysis helps identify the specific failure mechanisms of models under different physical artifacts, moving beyond simple accuracy metrics to understand structural and semantic fidelity.

Scanning & Warping: Impact of Geometric Deformation on Structure. The Scanning scenario (Table 2) reflects the ideal performance bound, where documents are flat and evenly lit. Specialized VLMs achieve character-level precision nearly identical to digital-born benchmarks. However, in the Warping scenario (Table 3), non-rigid deformations introduce complex curvilinear distortions that interfere with coordinate-based parsing. Our analysis indicates that while general VLMs maintain high recognition rates, their structural fidelity (TEDS) often declines more noticeably than that of PaddleOCR-VL-1.6. This suggests that paper curvature alters the spatial relationship between text lines, whereas models enhanced for geometric robustness can better anchor cell boundaries and formula components within distorted visual inputs.

Table 2: Sub-task: **Scanning**

Methods	Overall↑	Text^{E} ↓	CDM↑	TEDS↑	RO^{E} ↓
PP-StructureV3 [6]	84.68	0.094	84.34	79.06	0.092
Marker-1.8.2 [19]	70.27	0.223	77.03	56.05	0.238
Qwen3-VL-235B [1]	89.43	0.059	89.01	85.19	0.066
Gemini-3 Pro [9]	89.47	0.071	88.16	87.37	0.078
MinerU2.5-Pro [23]	92.11	<u>0.040</u>	89.77	90.57	**0.043**
GLM-OCR [27]	<u>92.67</u>	0.054	<u>91.10</u>	<u>92.28</u>	<u>0.061</u>
PaddleOCR-VL-1.6 [28]	**94.74**	**0.036**	**93.65**	**94.12**	**0.043**

Table 3: Sub-task: **Warping**

Methods	Overall↑	Text^{E} ↓	CDM↑	TEDS↑	RO^{E} ↓
PP-StructureV3 [6]	59.34	0.376	68.22	47.40	0.261
Marker-1.8.2 [19]	58.98	0.349	72.71	39.08	0.390
Qwen3-VL-235B [1]	89.99	<u>0.051</u>	89.06	85.95	<u>0.064</u>
Gemini-3 Pro [9]	88.90	0.086	88.10	87.20	0.087
MinerU2.5-Pro [23]	88.72	0.100	87.81	88.36	0.076
GLM-OCR [27]	<u>90.68</u>	0.071	<u>90.30</u>	<u>88.78</u>	0.100
PaddleOCR-VL-1.6 [28]	**92.48**	**0.049**	**91.63**	**90.66**	**0.061**

Screen-Photography & Illumination: Optical Artifacts and Consistency. Optical interference remains a significant bottleneck for real-world document intelligence. Moiré patterns in Screen-Photography (Table 4) and localized overexposure in Illumination (Table 5) often trigger "layout fragmentation" in traditional systems. Our experiments show that pipeline-based tools frequently misinterpret reflections as structural boundaries, leading to a sharp increase in Reading Order (RO^E) scores. In contrast, PaddleOCR-VL-1.6 demonstrates superior visual consistency. By integrating diverse lighting-augmented data during multi-task training, the model learns to utilize global contextual cues to maintain parsing precision even when local visual contrast is compromised.

Table 4: Sub-task: **Screen-Photo**

Methods	Overall↑	$Text^E$ ↓	CDM↑	TEDS↑	RO^E ↓
PP-StructureV3 [6]	66.89	0.204	73.26	47.82	0.165
Marker-1.8.2 [19]	63.65	0.290	72.73	47.21	0.325
Qwen3-VL-235B [1]	89.27	0.068	88.72	85.85	0.071
Gemini-3 Pro [9]	88.86	0.084	87.33	87.65	0.087
MinerU2.5-Pro [23]	91.29	0.050	87.41	91.44	**0.044**
GLM-OCR [27]	91.75	0.063	89.83	91.66	0.071
PaddleOCR-VL-1.6 [28]	**92.78**	**0.045**	**90.64**	**92.19**	0.054

Table 5: Sub-task: **Illumination**

Methods	Overall↑	$Text^E$ ↓	CDM↑	TEDS↑	RO^E ↓
PP-StructureV3 [6]	73.38	0.158	77.75	58.19	0.126
Marker-1.8.2 [19]	66.31	0.259	74.80	50.03	0.337
Qwen3-VL-235B [1]	89.27	0.060	87.81	86.05	0.070
Gemini-3 Pro [9]	89.53	0.073	87.78	88.14	0.080
MinerU2.5-Pro [23]	91.31	0.050	88.15	90.76	0.053
GLM-OCR [27]	91.12	0.059	91.02	88.20	0.071
PaddleOCR-VL-1.6 [28]	**93.28**	**0.042**	**92.61**	**91.48**	**0.051**

Skew: Perspective Correction and Reading Order Recovery. The Skew scenario (Table 6) represents one of the most demanding tasks in our benchmark, testing the limits of 3D-to-2D projection recovery. Large-angle perspective tilting induces a "line compression" effect, where text lines that are originally parallel appear to converge or overlap. Traditional systems relying on axis-aligned projections are prone to severe sequence errors. Experimental data show that PaddleOCR-VL-1.6 maintains a robust RO^E of 0.058, outperforming several large-scale general models. This performance suggests the model has internalized a flexible mapping of perspective cues, allowing it to accurately reconstruct the logical flow of a document even when the visual layout is aggressively skewed.

6 Discussion

Insights on Domain Gap and Robustness. Systematic evaluation on Real5-OmniDocBench reveals a pronounced performance degradation when models transition from digital-born environments to physical-world captures. Results indicate that while large-scale General VLMs possess strong semantic priors, their zero-shot transferability to non-rigid geometric transformations (*e.g.*, Warping) is not strictly correlated with parameter scale. This suggests the current bottleneck lies less in linguistic modeling and more in visual feature alignment under stochastic distortions. The persistence of this "reality gap" emphasizes the need for benchmarks quantifying robustness beyond standard accuracy, focusing on the intersection of 3D vision and document analysis.

Table 6: Sub-task: **Skew**

Methods	Overall↑	Text^{E} ↓	CDM↑	TEDS↑	RO^{E} ↓
PP-StructureV3 [6]	37.98	0.557	44.37	25.27	0.417
Marker-1.8.2 [19]	41.27	0.536	60.16	17.23	0.543
Qwen3-VL-235B [1]	86.56	<u>0.077</u>	83.96	83.41	<u>0.091</u>
Gemini-3 Pro [9]	<u>89.45</u>	0.080	<u>88.33</u>	<u>88.06</u>	0.092
MinerU2.5-Pro [23]	81.26	0.212	83.92	81.07	0.107
GLM-OCR [27]	85.39	0.099	85.78	80.28	0.156
PaddleOCR-VL-1.6 [28]	**92.66**	**0.045**	**91.44**	**91.04**	**0.058**

Efficacy of Specialized vs. General Architectures. Our results provide a nuanced perspective on architectural choices. Although General VLMs demonstrate superior adaptability in open-domain coherence, specialized models optimized for multi-task parsing often exhibit higher structural fidelity in constrained resource settings. The competitive performance of compact architectures suggests that domain-specific fine-tuning on high-quality augmented data can effectively compensate for smaller parameter counts. Conversely, failures in traditional Pipeline Tools under Skew and Illumination highlight the limitations of heuristic systems, advocating for unified end-to-end architectures that jointly optimize optical correction and structural parsing.

7 Conclusion

In this paper, we introduced Real5-OmniDocBench, a comprehensive benchmark evaluating document parsing robustness across five real-world scenarios. By reconstructing 1,355 high-fidelity images based on the OmniDocBench v1.5 corpus, we provide a standardized platform for assessing how physical artifacts impact model performance.

Our benchmarking reveals significant vulnerabilities in current parsing technologies, particularly regarding complex 3D distortions and localized illumination variations. Results highlight that while compact, specialized models achieve high efficiency, there remains a substantial margin for universal robustness. We hope Real5-OmniDocBench will catalyze research toward more resilient models capable of handling the inherent complexities of practical applications.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D.,

Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631

2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
3. Chen, S., Guo, X., Li, Y., Zhang, T., Lin, M., Kuang, D., Zhang, Y., Ming, L., Zhang, F., Wang, Y., Xu, J., Zhou, Z., Chen, W.: Ocean-ocr: Towards general ocr application via a vision-language model (2025), https://arxiv.org/abs/2501.15558
4. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., Zhang, Y., Zhang, Y., Liu, Y., Yu, D., Ma, Y.: Paddleocr-vl-1.5: Towards a multi-task 0.9b vlm for robust in-the-wild document parsing (2026), https://arxiv.org/abs/2601.21957
5. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., Zhang, Y., Zhang, Y., Zheng, H., Zhang, J., Zhang, J., Liu, Y., Yu, D., Ma, Y.: Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model (2025), https://arxiv.org/abs/2510.14528
6. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., Zhang, Y., Lv, W., Huang, K., Zhang, Y., Zhang, J., Zhang, J., Liu, Y., Yu, D., Ma, Y.: Paddleocr 3.0 technical report (2025), https://arxiv.org/abs/2507.05595
7. Du, Y., Chen, P., Ying, X., Chen, Z.: Docptbench: Benchmarking end-to-end photographed document parsing and translation (2025), https://arxiv.org/abs/2511.18434
8. Google DeepMind: Gemini 2.5. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ (2025)
9. Google DeepMind: Gemini 3.0. https://blog.google/products-and-platforms/products/gemini/gemini-3-collection/ (2025)
10. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019)
11. Li, G., Lyu, P., Zhang, C., Shen, H., Wu, L., Wan, X., Zeng, G., Hu, H., Ma, C., Zhou, Y.: Towards real-world document parsing via realistic scene synthesis and document-aware training (2026), https://arxiv.org/abs/2603.23885
12. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots.ocr: Multilingual document layout parsing in a single vision-language model (2025), https://arxiv.org/abs/2512.02498
13. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Yang, B., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm (2026), https://arxiv.org/abs/2506.05218
14. Mandal, S., Talewar, A., Ahuja, P., Juvatkar, P.: Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging (2025)
15. Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., Jin, Z., Liang, G., Zhang, R., Zhang, W., Qu, Y., Ren, Z., Sun,

Y., Zheng, Y., Ma, D., Tang, Z., Niu, B., Miao, Z., Dong, H., Qian, S., Zhang, J., Chen, J., Wang, F., Zhao, X., Wei, L., Li, W., Wang, S., Xu, R., Cao, Y., Chen, L., Wu, Q., Gu, H., Lu, L., Wang, K., Lin, D., Shen, G., Zhou, X., Zhang, L., Zang, Y., Dong, X., Wang, J., Zhang, B., Bai, L., Chu, P., Li, W., Wu, J., Wu, L., Li, Z., Wang, G., Tu, Z., Xu, C., Chen, K., Qiao, Y., Zhou, B., Lin, D., Zhang, W., He, C.: Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing (2025), https://arxiv.org/abs/2509.22186

16. OpenAI.: Gpt-5.2 system card (2025), https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf
17. opendatalab: Mineru2.0-2505-0.9b. https://huggingface.co/opendatalab/MinerU2.0-2505-0.9B (2025)
18. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., Shi, J., Wu, F., Chu, P., Liu, M., Li, Z., Xu, C., Zhang, B., Shi, B., Tu, Z., He, C.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations (2024), https://arxiv.org/abs/2412.07626
19. Paruchuri, V.: Marker. https://github.com/datalab-to/marker (2025), accessed: 2025-09-25
20. Poznanski, J., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Rangapur, A., Wilhelm, C., Lo, K., Soldaini, L.: olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models (2025), https://arxiv.org/abs/2502.18443
21. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S.H., Cao, Y., Charles, Y., Che, H.S., Chen, C., Chen, G., Chen, H., Chen, J., Chen, J., Chen, J., Chen, J., Chen, K., Chen, L., Chen, R., Chen, X., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Z., Chen, Z., Cheng, D., Chu, M., Cui, J., Deng, J., Diao, M., Ding, H., Dong, M., Dong, M., Dong, Y., Dong, Y., Du, A., Du, C., Du, D., Du, L., Du, Y., Fan, Y., Fang, S., Feng, Q., Feng, Y., Fu, G., Fu, K., Gao, H., Gao, T., Ge, Y., Geng, S., Gong, C., Gong, X., Gongque, Z., Gu, Q., Gu, X., Gu, Y., Guan, L., Guo, Y., Hao, X., He, W., He, W., He, Y., Hong, C., Hu, H., Hu, J., Hu, Y., Hu, Z., Huang, K., Huang, R., Huang, W., Huang, Z., Jiang, T., Jiang, Z., Jin, X., Jing, Y., Lai, G., Li, A., Li, C., Li, C., Li, F., Li, G., Li, G., Li, H., Li, H., Li, J., Li, J., Li, J., Li, L., Li, M., Li, W., Li, W., Li, X., Li, X., Li, Y., Li, Y., Li, Y., Li, Y., Li, Z., Li, Z., Liao, W., Lin, J., Lin, X., Lin, Z., Lin, Z., Liu, C., Liu, C., Liu, H., Liu, L., Liu, S., Liu, S., Liu, S., Liu, T., Liu, T., Liu, W., Liu, X., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Z., Liu, Z., Lu, E., Lu, H., Lu, Z., Luo, J., Luo, T., Luo, Y., Ma, L., Ma, Y., Mao, S., Mei, Y., Men, X., Meng, F., Meng, Z., Miao, Y., Ni, M., Ouyang, K., Pan, S., Pang, B., Qian, Y., Qin, R., Qin, Z., Qiu, J., Qu, B., Shang, Z., Shao, Y., Shen, T., Shen, Z., Shi, J., Shi, L., Shi, S., Song, F., Song, P., Song, T., Song, X., Su, H., Su, J., Su, Z., Sui, L., Sun, J., Sun, J., Sun, T., Sung, F., Tai, Y., Tang, C., Tang, H., Tang, X., Tang, Z., Tao, J., Teng, S., Tian, C., Tian, P., Wang, A., Wang, B., Wang, C., Wang, C., Wang, C., Wang, D., Wang, D., Wang, D., Wang, F., Wang, H., Wang, H., Wang, H., Wang, H., Wang, H., Wang, J., Wang, J., Wang, J., Wang, K., Wang, L., Wang, Q., Wang, S., Wang, S., Wang, S., Wang, W., Wang, X., Wang, X., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wang, Z., Wang, Z., Wang, Z., Wang, Z., Wang, Z., Wei, C., Wei, M., Wen, C., Wen, Z., Wu, C., Wu, H., Wu, J., Wu, R., Wu, W., Wu, Y., Wu, Y., Wu, Y., Wu, Z., Xiao, C., Xie, J., Xie, X., Xie, Y., Xin, Y., Xing, B., Xu, B., Xu, J., Xu, J., Xu, J., Xu, L.H., Xu, L., Xu, S., Xu, W., Xu, X., Xu, X., Xu, Y., Xu, Y., Xu, Y., Xu, Z., Xu, Z., Yan, J., Yan, Y., Yang, G., Yang, H., Yang, J., Yang, K., Yang, N., Yang, R., Yang, X., Yang, X., Yang, Y., Yang, Y., Yang, Y., Yang, Z., Yang,

Z., Yang, Z., Yao, H., Ye, D., Ye, W., Ye, Z., Yin, B., Yu, C., Yu, L., Yu, T., Yu, T., Yuan, E., Yuan, M., Yuan, X., Yue, Y., Zeng, W., Zha, D., Zhan, H., Zhang, D., Zhang, H., Zhang, J., Zhang, P., Zhang, Q., Zhang, R., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, C., Zhao, F., Zhao, J., Zhao, S., Zhao, X., Zhao, Y., Zhao, Z., Zheng, H., Zheng, R., Zheng, S., Zheng, T., Zhong, J., Zhong, L., Zhong, W., Zhou, M., Zhou, R., Zhou, X., Zhou, Z., Zhu, J., Zhu, L., Zhu, X., Zhu, Y., Zhu, Z., Zhuang, J., Zhuang, W., Zou, Y., Zu, X.: Kimi k2.5: Visual agentic intelligence (2026), https://arxiv.org/abs/2602.02276

22. Wang, A.L., Tang, J., Lei, L., Feng, H., Liu, Q., Fei, X., Lu, J., Wang, H., Liu, W., Liu, H., Liu, Y., Bai, X., Huang, C.: Wilddoc: How far are we from achieving comprehensive and robust document understanding in the wild? (2025), https://arxiv.org/abs/2505.11015
23. Wang, B., He, T., Ouyang, L., Wu, F., Zhao, Z., Chu, T., Qu, Y., Jin, Z., Zeng, W., Miao, Z., et al.: Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. arXiv preprint arXiv:2604.04771 (2026)
24. Wang, B., Wu, F., Ouyang, L., Gu, Z., Zhang, R., Xia, R., Shi, B., Zhang, B., He, C.: Image over text: Transforming formula recognition evaluation with character detection matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19681–19690 (2025)
25. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression (2025), https://arxiv.org/abs/2510.18234
26. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr 2: Visual causal flow (2026), https://arxiv.org/abs/2601.20552
27. zai-org: Glm-ocr. https://huggingface.co/zai-org/GLM-OCR (2026)
28. Zhang, Z., Liu, H., Liang, S., Zhang, Y., Xiang, Y., Liu, J., Sun, T., Lin, M., Zhang, Y., Zhou, C., Gao, T., Cui, C., Liu, Y., Yu, D., Ma, Y.: Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training (2026), https://arxiv.org/abs/2606.03264
29. Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: European conference on computer vision. pp. 564–580. Springer (2020)