SIMDOCBENCH: BENCHMARKING THE ROBUSTNESS OF DOCUMENT PARSING VIA PAPER REALISTIC SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern optical character recognition (OCR) and document parsing systems convert document images into machine-readable text and structured content, and now achieve strong performance on clean, scanned documents. However, their robustness to real-world physical capture remains poorly understood, as photographed documents often exhibit diverse geometric, photometric, and environmental variations that are not systematically evaluated by existing benchmarks. To address this gap, we introduce SimDocBench with a *Source–Neutral–Perturbed protocol* that separates rendering effects from acquisition-induced degradation. Using physically based thin-sheet simulation, path-traced rendering, and camera-effect modeling, SimDocBench constructs 158,496 single-factor renders covering 31 acquisition factors across 5 dimensions, together with *Regular* and *Hard* multi-factor settings. Rather than measuring parsing accuracy alone, SimDocBench evaluates three complementary properties of document parsers: accuracy under physical acquisition changes, consistency across content-preserving observations of the same document, and recoverability through controlled input restoration. Extensive experiments reveal that physical robustness is not simply an extension of clean-document accuracy: models with similar clean performance can exhibit substantially different degradation, particularly under challenging acquisition conditions. Moreover, accuracy degradation and prediction consistency capture distinct robustness behaviors, while controlled restoration further reveals heterogeneous failure mechanisms across acquisition factors and parsers. Together, these results show that physical robustness is a distinct and multi-dimensional capability largely invisible to clean-document benchmarks.

1 INTRODUCTION

Document parsing aims to recover text, equations, tables, and reading order from document images, serving as a fundamental component of document understanding and downstream knowledge-centric applications. With the rapid development of vision-language models and end-to-end parsing systems (Team et al., 2025; Cui et al., 2026; Wu et al., 2026; Wang et al., 2026), recent systems have achieved strong performance on clean digital and scanned documents, while evaluation has expanded from text recognition to full structured document parsing (Ouyang et al., 2025).

However, real-world document images are often acquired under substantially different conditions from clean digital or scanned inputs. In many practical scenarios, documents are captured directly with mobile phones or cameras, introducing variations in viewpoint, illumination, paper geometry, background, and imaging quality. Such physical acquisition changes can substantially alter the observed document image even when its underlying content remains unchanged, posing additional challenges to document parsing (Wang et al., 2025; Du et al., 2025).

Recent benchmarks have begun to evaluate document parsing under real-world acquisition conditions, but they remain limited in disentangling the effects of individual factors (Wang et al., 2025; Du et al., 2025). During physical capture, camera pose, illumination, paper deformation, background, and imaging effects often vary simultaneously, making it difficult to isolate and reproduce a specific acquisition change. Consequently, performance differences across capture conditions cannot be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

readily attributed to individual factors (Zhou et al., 2026). This raises a fundamental question: *how do different physical acquisition factors affect document parsing when the underlying document content is preserved?*

To address these limitations, we introduce **SimDocBench**, a benchmark for systematically evaluating full document parsing under controlled physical acquisition changes. We combine physically based thin-sheet paper simulation (Narain et al., 2012), path-traced rendering (Kajiya, 1986), and camera-effect modeling to generate realistic and controllable document images while preserving annotation correspondence through deformation and projection. Built on clean documents, SimDocBench contains 158,496 single-factor renders covering 31 acquisition factors across 5 dimensions: camera imaging, illumination, paper geometry, surface, and background, together with *SimDocBench-Regular* and *SimDocBench-Hard* for multi-factor evaluation. This design enables matched comparisons of the same document under controlled single- and multi-factor acquisition changes.

Rather than measuring parsing accuracy alone, SimDocBench evaluates document parsers from three complementary perspectives under a Source–Neutral–Perturbed protocol. First, *accuracy under physical acquisition changes* separates the source-to-neutral rendering gap from the neutral-to-perturbed acquisition gap, with matched neutral controls supporting factor-level analysis. Second, *prediction consistency* examines whether a parser produces consistent structured outputs across content-preserving observations of the same document. Third, *recoverability* measures how much lost parsing performance can be recovered through controlled input restoration, such as geometry or background correction.

Extensive experiments on eight representative document parsers reveal that clean-document accuracy is not a reliable proxy for physical robustness: models with similar clean performance can respond very differently to acquisition changes, with substantially larger degradation under the Hard setting. Moreover, accuracy degradation and prediction consistency capture distinct robustness behaviors, while controlled restoration reveals factor- and model-specific failure mechanisms. These findings show that physical robustness is a distinct and multi-dimensional capability that is largely overlooked by clean-document evaluation. Our main contributions are as follows:

- We highlight the importance of systematically evaluating document parsing under real-world physical acquisition changes, and introduce a Source–Neutral–Perturbed protocol that separates rendering effects from acquisition-induced degradation.
- We introduce SimDocBench, a controllable benchmark with 158,496 single-factor renders covering 31 acquisition factors across five dimensions, together with Regular and Hard multi-factor settings. Its physics-based simulation and rendering pipeline enables realistic and independently controlled acquisition changes.
- We systematically evaluate document parsers from three complementary perspectives: accuracy under physical acquisition changes, prediction consistency, and recoverability through controlled restoration. The results reveal physical robustness behaviors that cannot be reliably inferred from clean-document performance alone.

2 Related Work

While document parsing models and benchmarks have advanced significantly in recent years, their robustness has received increasing attention due to the challenges posed by diverse acquisition conditions and real-world document variations. We review robustness in document parsing, particularly physical document transformations and controlled evaluation protocols.

2.1 Document Parsing and Capture Robustness

Document benchmarks have expanded from mobile capture and recognition (Burie et al., 2015) to broader OCR capabilities and structured parsing. OCRBench and OCRBench v2 cover recognition, localization, and visual-text reasoning (Liu et al., 2024; Fu et al., 2025), while OmniDocBench evaluates text, formulas, tables, and reading order across diverse PDF sources (Ouyang et al., 2025). Other evaluations examine downstream consequences of OCR errors in retrieval-augmented generation (Sun et al., 2026). More recently, WildDoc, DocPTBench, and Wild-OmniDocBench have extended evaluation to photographed documents, exposing substantial gaps between digital inputs

and real capture (Wang et al., 2025; Du et al., 2025; Li et al., 2026a). MDPBench adds multilingual coverage (Li et al., 2026b), and PureDocBench provides source-traceable clean and degraded versions, showing that degradation can reverse model rankings (Li et al., 2026c). These findings have motivated deformation-aware learning and diagnostic-guided data construction (Cai et al., 2026; Yu et al., 2026). Real5-OmniDocBench further establishes one-to-one physical correspondence across scanning, warping, screen-photography, illumination, and skew, supporting comparisons across capture scenarios (Zhou et al., 2026).

2.2 Physical Document Modeling and Restoration

Synthetic document modeling has long supported geometric restoration. DocUNet generates paired data with 2D mesh warps (Ma et al., 2018); DewarpNet/Doc3D and UVDoc provide three-dimensional document geometry and surface-to-image correspondence for learning rectification (Das et al., 2019; Verhoeven et al., 2023). Rendered documents also support layout learning through domain randomization, which varies synthetic scene appearance to expose models to diverse visual conditions (Ling et al., 2021). Subsequent restoration methods exploit document boundaries, text lines, layout structure, and progressive correction (Jiang et al., 2022; Feng et al., 2022; Li et al., 2024; Feng et al., 2025). Beyond geometry, intrinsic decomposition separates reflectance and shading, while methods such as DocTr perform geometric and illumination correction and evaluate their benefits for OCR (Das et al., 2020; Feng et al., 2021), establishing that restoration quality should also be assessed through recognition performance.

2.3 Controlled Robustness and Metamorphic Evaluation

ImageNet-C/P established standardized evaluation of robustness to image perturbations and prediction stability (Hendrycks & Dietterich, 2019), while subsequent work showed that measured robustness can depend on similarity between training augmentations and test perturbations (Mintun et al., 2021). This approach extends to multimodal capabilities in MLLM-IC (Qiu et al., 2025) and structured vision in RobustSpring, which preserves temporal, stereo, and depth consistency under image perturbations (Schmalfuss et al., 2025). Document-focused diagnostics additionally examine naturally occurring acquisition properties and controlled perturbations of visual-text reasoning inputs (Peng et al., 2026; Cheng et al., 2026). Metamorphic testing provides complementary checks by specifying expected relations between predictions across transformed inputs (Tian et al., 2018; Dwarakanath et al., 2018).

SimDocBench targets full document parsing by combining physical simulation and rendering to synthesize document images under simulated in-the-wild acquisition conditions while preserving correspondence with source-page annotations. By varying individual factors while holding document content and other acquisition conditions fixed, it enables controlled analysis of how these acquisition factors affect the parsing of text, formulas, tables, and reading order.

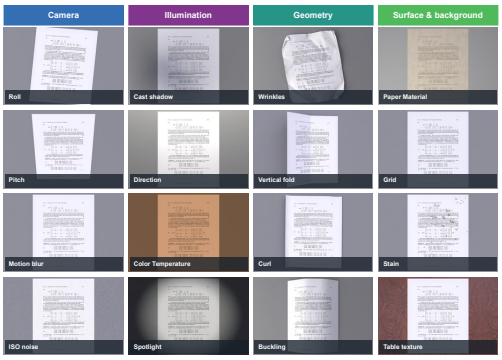
3 The SimDocBench Dataset and Benchmark

In this section, we define the document parsing task and acquisition factors, describe the source documents and construction pipeline, and summarize the dataset composition.

3.1 Acquisition Factor Taxonomy

Given a page image, a parser m must produce a structured representation of its text, equations and tables. For a source page p rendered under acquisition condition c, we write $(X_{p,c}, Y_{p,c})$ for the rendered image and its propagated annotation; the task is to recover the structure in $Y_{p,c}$ from $X_{p,c}$.

For acquisition conditions encountered in the real world, we group the 31 acquisition factors into five categories, summarized in Figure 1. Camera factors vary viewpoint and imaging, while Illumination factors change how light reaches the page. Paper Geometry changes the three-dimensional surface, potentially coupling projected layout with shading. Surface factors alter paper material or add visual distractors on the page, whereas Background varies the supporting surface rather than the page itself.



(a) Representative visual examples.

Camera (11): Pitch, yaw, roll; focal length, exposure, ISO noise; blur magnitude and direction; lens distortion, chromatic aberration, vignetting.

Illumination (4): Light direction, color temperature, shadow, spotlight.

Paper geometry (11): Concave, convex, compound bends; horizontal and vertical folds; curl, dog ear, crease, wrinkles, twist, buckling.

Surface (4): Paper material, grid, splatter, watermark.

Background (1): Table texture.

(b) Taxonomy of acquisition factors.

Figure 1: Overview of SimDocBench acquisition factors. (a) Representative perturbations of one document. (b) The complete taxonomy of 31 acquisition factors in five groups.

Each acquisition factor f is associated with a predefined set of candidate configurations. For a source page p and an acquisition factor f, let $V_{f,p}$ denote the set of variant indices sampled for factor f on page p, where each $v \in V_{f,p}$ corresponds to a specific sampled configuration. For factors with an ordering in severity, such as roll, pitch and motion blur, we sample three candidate configurations per page at different intensity levels. For example, for motion blur, we use blur magnitudes of 0.75, 1.5, and 3 pixels. We do not consider stronger blur levels, as manual inspection indicates that motion blur beyond 3 pixels causes irreversible degradation of textual content, making the resulting images unsuitable for meaningful recovery or recognition. For categorical acquisition factors, we instead construct a candidate configuration pool for each factor, such as watermark, stains, and background textures. For paper material, we consider six configurations: office matte paper, coated glossy paper, textured parchment, rough recycled paper, high-gloss photo paper, and newsprint. Candidate settings and sampling rules are detailed in Appendix B.1.

3.2 Data Sourcing and Construction

We use 1,651 annotated pages from OmniDocBench (Ouyang et al., 2025) as source documents, comprising 1,355 pages from the v1.5 subset and 296 additional pages emphasizing formulas, layouts, and tables. Building on this collection, we construct our dataset through the following pipeline.

Paper Simulation. We first map each source image onto a triangular paper mesh using UV coordinates. We then use ARCSim (Adaptive Refining and Coarsening Simulator) thin-sheet simulation (Narain et al., 2012; 2013) to model paper deformation, including bending, stretching, and contact with other surfaces. ARCSim provides a physically based model for thin-sheet deformation, allowing us to simulate realistic paper folding behaviors while preserving geometric consistency. This helps generate plausible deformations that are difficult to obtain with manually designed transformations. The deformed mesh keeps the original UV information, allowing the page texture and annotations to move together with the paper. To improve efficiency, meshes are generated in advance for different page aspect ratios and reused when needed. More details about the mesh construction are provided in Appendix B.1.

Rendering and Camera Effects. Rendering connects the simulated paper shape with the scene lighting and camera view. We use the original page as a UV texture on the deformed mesh and place it above a supporting surface in Blender. For each capture setting, we define the paper and background materials, light sources, and camera parameters. Cycles uses physically based path tracing (Kajiya, 1986; Blender Foundation, n.d.) to simulate how light interacts with the scene. As a result, a fold can change the projected text shape, visibility, and shading at the same time. The paper and the supporting surface can also create reflections and shadows depending on their materials. After rendering, we add realistic camera effects in postprocessing, including motion blur from small image shifts, lens distortion, chromatic aberration, vignetting, and sensor noise controlled by an ISO-related parameter.

Annotation Propagation. Let the width and height of the source page be W_p and H_p, respectively, with image coordinates defined relative to the top-left corner. For each source point (x, y), we first convert it to UV coordinates as $(u, v) = (x/W_p, 1 - y/H_p)$. We then recover its corresponding 3D position $q \in \mathbb{R}^3$ via barycentric interpolation of the three vertices of the triangle containing the UV coordinate. Finally, the 3D point is projected into the image plane using the projection matrix of the camera. Through this procedure, annotations on the source page p, including text locations and bounding-box coordinates, are mapped onto the rendered image $X_{p,c}$, yielding the propagated annotation $Y_{p,c}$.

Visibility and Readability Validation. We screen rendered images for annotated-region orientation and self-occlusion, field-of-view coverage, text scale, and brightness, adjusting resolution for small text and discarding unusable renders. Appendix B.1 details the screening criteria.

3.3 Dataset Statistics

The 1,651 sources comprise 1,355 OmniDocBench v1.5 pages and 296 targeted formula, layout, or table pages; they yield 158,496 single-factor renders. The two multi-factor settings, SimDocBench-Regular and SimDocBench-Hard, share the same 1,651 source pages. Regular sampling draws five compatible views per source (8,255 images). Hard assigns one compound-stress view per source, combining physical wrinkles, rough recycled paper, camera roll and pitch, splatter and table texture, plus optional grid and blur. Pose-aware projection permits wrinkle–pose combinations excluded from regular sampling. Visibility-approved retries and lower-intensity fallbacks make the Hard setting a mixed-severity collection. Appendix B.1 details the settings, selection, and mixed-strength composition.

4 SimDocBench Evaluation Methodology

4.1 Evaluation Framework

Figure 2 organizes three comparisons around neutral renders. Even without an acquisition perturbation, rendering can change image resolution and brightness. We therefore render each source page under baseline conditions as a canonical neutral reference f_0: flat paper on a neutral-gray background, viewed head-on under uniform illumination and proper exposure.

The source–neutral pair reveals the baseline change introduced by rendering. For single-factor sensitivity, we compare each factor's sampled variants with its assigned neutral control N_f on the same page; this factor-matched control may differ from the canonical reference because rendering recipes vary. For multi-factor robustness, we compare combined perturbations with the canonical neutral reference. SimDocBench-Regular samples compatible factor combinations, while SimDocBench-Hard stresses particularly challenging conditions; both use the same source pages. The following subsection defines the corresponding scores and gaps.

We use the OmniDocBench evaluator (Ouyang et al., 2025) for output normalization, element matching, and parsing accuracy, using text edit distance, formula CDM, and table TEDS with the remapped ground truth for each rendered image. We follow its task-balanced aggregation for the overall score $A_m(c)$ of parser m under condition c. Let $z_{m,p,c} = F_m(X_{p,c})$ denote the prediction of parser m for source page p under condition c.

4.2 Paired Evaluation of Acquisition Perturbations

Source–Neutral–Perturbed Comparison. For aggregate evaluation, we compare source images S, the canonical neutral reference f_0, and regular multi-factor perturbations M on the same source pages. We measure

$$\Delta_m^{\text{render}} = A_m(S) - A_m(f_0), \qquad \Delta_m^{\text{acq}} = A_m(f_0) - A_m(M). \tag{1}$$

The rendering gap measures the score change from the source image to its neutral rendering. The acquisition gap measures the additional score change under the sampled acquisition perturbations relative to that baseline. Their sum equals the Source–Perturbed gap, making the contribution of each

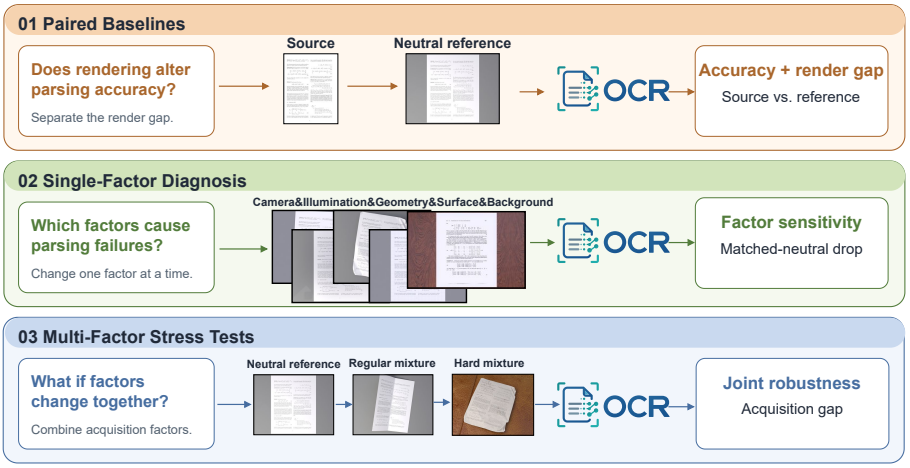


Figure 2: Overview of SimDocBench's three evaluation settings: paired baselines for the rendering gap, single-factor sensitivity for diagnosis, and multi-factor scenes for robustness.

comparison explicit. Both are measured in score points: positive values indicate degradation, and negative values indicate improvement. We report absolute scores alongside gaps to characterize both parsing accuracy and sensitivity to acquisition changes.

For single-factor analysis, we compare sampled variants $P_{f,v}$ with the factor-matched neutral control N_f on the same source-page support:

$$\mathrm{PertDrop}_{m,f} = A_m(N_f) - A_m(P_f). \tag{2}$$

Here $A_m(P_f)$ uses the same three-task aggregation as the main result: average sampled variants per source page within each task, aggregate text edit accuracy, formula CDM, and raw table TEDS on their respective task supports, and give the three tasks equal weight. Paired source-page bootstrap replicates estimate uncertainty in the aggregate difference; no per-page overall score is constructed.

4.3 Prediction Consistency

We pair accuracy gaps with structure-aware metamorphic consistency to measure both degradation and output stability. A metamorphic relation specifies how outputs should relate after an input transformation (Chen et al., 2020). Here, acquisition perturbations should preserve the parsed content and structure when the document remains visible and legible. We evaluate two relations on the same source page: rendering consistency (MR1) compares predictions for the original source image S and its canonical neutral render f_0, while acquisition consistency (MR2) compares predictions for a neutral render N_f and a single-factor perturbation $P_{f,v}$. Our consistency analysis directly compares paired predictions, following the distinction between accuracy and stability in perturbation benchmarks (Schmalfuss et al., 2025).

For paired predictions $z = z_{m,p,c}$ and $z' = z_{m,p,c'}$, we introduce a benchmark-specific Metamorphic Consistency Distance (MCD):

$$\mathrm{MCD}(z, z') = \frac{1}{|\mathcal{C}(z, z')|} \sum_{k \in \mathcal{C}(z,z')} d_k(z, z'), \tag{3}$$

where $\mathcal{C}$ contains applicable text, formula, table, and order components. Each distance is one minus the corresponding similarity, so lower MCD indicates greater prediction stability.

5 Experiments

We report overall accuracy under the Regular and Hard settings, then examine which individual factors expose failures. Additional setup, evaluation details, and full results are provided in Appendices B–D.

Table 1: Task-balanced parsing accuracy under Regular and Hard acquisition. $\Delta_R = S - N$, $\Delta_M = N - M$, and $\Delta_H = N_H - H$; N_H is the Hard matched neutral. Hard text is edit accuracy, formula is CDM, table is TEDS, and order is separate.

	Regular					Hard			Hard components			
System	S	N	M	Δ_R	Δ_M	N_H	H	Δ_H	Text	Form.	Table	Order
PaddleOCR-VL-1.6	96.12	95.04	92.85	1.08	2.19	95.04	78.61	16.44	79.99	81.59	74.24	72.37
OvisOCR2	95.81	95.04	92.27	0.76	2.78	94.28	76.41	17.87	81.82	84.25	63.15	76.86
MinerU2.5-Pro	95.66	94.48	91.61	1.19	2.87	94.48	54.19	40.29	45.96	65.77	50.83	57.74
GLM-OCR	95.51	93.43	91.52	2.07	1.91	93.43	72.48	20.95	69.82	80.89	66.72	66.47
PaddleOCR-VL-1.5	94.86	93.07	90.82	1.80	2.25	93.07	78.06	15.01	79.76	80.71	73.71	72.24
FireRed-OCR	93.02	89.60	87.60	3.43	1.99	88.48	68.72	19.76	75.24	77.25	53.67	70.45
Qianfan-OCR	93.89	87.99	84.17	5.90	3.82	87.37	50.41	36.95	54.53	54.46	42.24	58.16
HunyuanOCR	89.24	86.53	83.53	2.71	3.00	86.35	66.64	19.71	73.31	72.40	54.22	70.57

5.1 Evaluation Details

Evaluated systems. We evaluate eight vision-language document parsing systems: PaddleOCR-VL-1.5 (Cui et al., 2026) and PaddleOCR-VL-1.6 (Zhang et al., 2026), OvisOCR2 (Lu et al., 2026), MinerU2.5-Pro (Wang et al., 2026), GLM-OCR (Duan et al., 2026), FireRed-OCR (Wu et al., 2026), Qianfan-OCR (Dong et al., 2026), and HunyuanOCR (Team et al., 2025).

PaddleOCR-VL-1.5, PaddleOCR-VL-1.6, and GLM-OCR use a layout detector (Sun et al., 2025) followed by region recognition. MinerU2.5-Pro performs coarse layout analysis and fine region recognition within one model. The other four systems parse the full page end to end. The eight-system cohort thus spans region-based and full-page parsing designs.

Inputs and scores. SimDocBench-Regular contains 1,651 source pages. For each page, we provide a neutral render and five sampled multi-factor views. SimDocBench-Hard contains one image for each of the same 1,651 source pages. Both source and neutral comparisons are evaluated using the corresponding source-page annotations. All accuracy analyses use the same task-balanced parsing score defined in Section 4.2; MCD separately measures prediction consistency. Single-factor and regular multi-factor scenes are rendered on a default 2048×1536 canvas with 32 Cycles samples. For dense pages, the resolution is adjusted following the procedure described in Section 3.2.

5.2 Overall Parsing Accuracy under Acquisition Changes

Accuracy lost under acquisition changes. Acquisition changes reduce parsing accuracy beyond the loss caused by neutral rendering. As Table 1 shows, neutral scores are 0.76–5.90 lower than source scores across the eight systems, and regular acquisition lowers them by a further 1.91–3.82. The strongest parser under regular acquisition is not the least affected by it. PaddleOCR-VL-1.6 leads with a regular score of 92.85, while GLM-OCR scores 91.52 but has the smallest acquisition gap.

Regular and Hard acquisition. Hard acquisition reveals substantially larger losses. Across all source pages, acquisition gaps range from 15.01 to 40.29; MinerU2.5-Pro, for example, scores 94.48 on the matched neutral renders but only 54.19 under Hard acquisition. Each gap compares matched source-page support. The rightmost columns of Table 1 break down Hard accuracy by task.

5.3 Factor-Level Sensitivity

Acquisition factors and degradation. Wrinkles cause the largest mean degradation in the official task-balanced score. Figure 3 shows an eight-system mean PertDrop of 4.89 against factor-matched neutral controls, followed by paper material at 3.27, camera roll at 2.91, vertical folds at 2.50, and grid underlay at 2.48. These factor-specific results reveal vulnerabilities hidden by the aggregate scores in Table 1. The same factor can affect parsers very differently. Under wrinkles, PertDrop ranges from 0.50 for PaddleOCR-VL-1.5 to 14.97 for Qianfan-OCR, with MinerU2.5-Pro at 9.56. Under camera roll, PaddleOCR-VL-1.5 has a PertDrop of -0.30, compared with 7.94 for Qianfan-OCR. The cast-shadow factor has a negative eight-system mean PertDrop of -1.49 under the sampled settings.

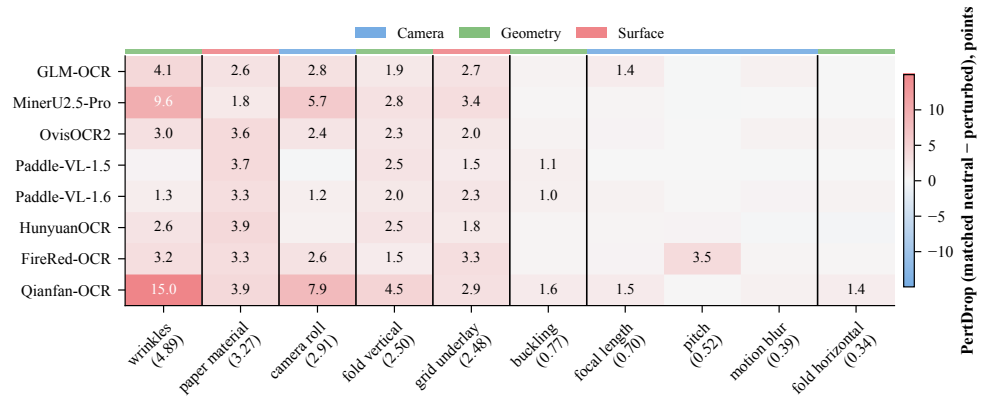


Figure 3: Ten factors with the largest eight-system mean PertDrop. Values below factor names give the eight-system mean in score points; the full heatmap can be found in Appendix 6.

6 Further Analysis

We examine whether accuracy and prediction consistency agree, how much accuracy can be restored by simple input corrections, and how simulated results transfer to real acquisition.

Accuracy and Prediction Consistency Figure 4 shows that neutral and regular accuracy rank the eight systems similarly (Spearman $\rho = 0.976$), but neutral accuracy does not closely track the acquisition gap ($\rho = -0.357$). Table 2 reports overall and component MCD distances: overall MCD spans 0.103–0.298 for source-to-neutral pairs (MR1) and 0.054–0.204 for neutral-to-acquisition pairs (MR2). Under MR2, MinerU2.5-Pro has a mean formula distance of 0.182 and a mean text distance of 0.034 on their respective applicable pairs; these are continuous distances, not thresholded violation rates. Lower MCD indicates more stable predictions within a relation; MR1 and MR2 use different pairs and should not be interpreted as a matched severity comparison. PertDrop and MCD are weakly associated across 176 model–factor cells ($\rho = 0.199$), and 316 of 616 within-factor model-pair orderings reverse. Accuracy, degradation, and prediction stability therefore capture different aspects of these systems.

Table 2: Mean overall MCD and component distances for MR1 (source–neutral) and MR2 (neutral–acquisition). Smaller values indicate more stable predictions.

	MR1					MR2				
System	MCD	Text	Form.	Table	Order	MCD	Text	Form.	Table	Order
MinerU2.5-Pro	0.128	0.051	0.374	0.089	0.135	0.075	0.034	0.182	0.050	0.086
GLM-OCR	0.170	0.080	0.372	0.071	0.209	0.103	0.052	0.190	0.039	0.131
PaddleOCR-VL-1.6	0.118	0.051	0.238	0.066	0.149	0.074	0.027	0.167	0.047	0.085
PaddleOCR-VL-1.5	0.133	0.057	0.277	0.077	0.168	0.077	0.028	0.178	0.064	0.092
OvisOCR2	0.105	0.067	0.188	0.056	0.124	0.054	0.036	0.087	0.031	0.066
HunyuanOCR	0.298	0.188	0.489	0.186	0.379	0.204	0.137	0.263	0.124	0.233
FireRed-OCR	0.103	0.054	0.189	0.063	0.132	0.072	0.040	0.125	0.042	0.091
Qianfan-OCR	0.163	0.110	0.217	0.099	0.196	0.127	0.096	0.154	0.079	0.138

Restoration Analysis To measure how much parsing accuracy can be recovered by simple input corrections, we apply renderer-guided geometry rectification and background replacement before parsing. We compare each perturbed image on a matched 150-page diagnostic subset (three variants per factor) with geometry-only rectification, background-only replacement, their combination, and the canonical neutral reference on the same 2048×1536 canvas. Geometry rectification uses renderer UV correspondences; background replacement uses the projected paper mask. Gains are differences from the original perturbed input in the official task-balanced score, with paired source-page bootstrap intervals (Appendix C.4). Under camera roll, geometry rectification improves the score by 7.97

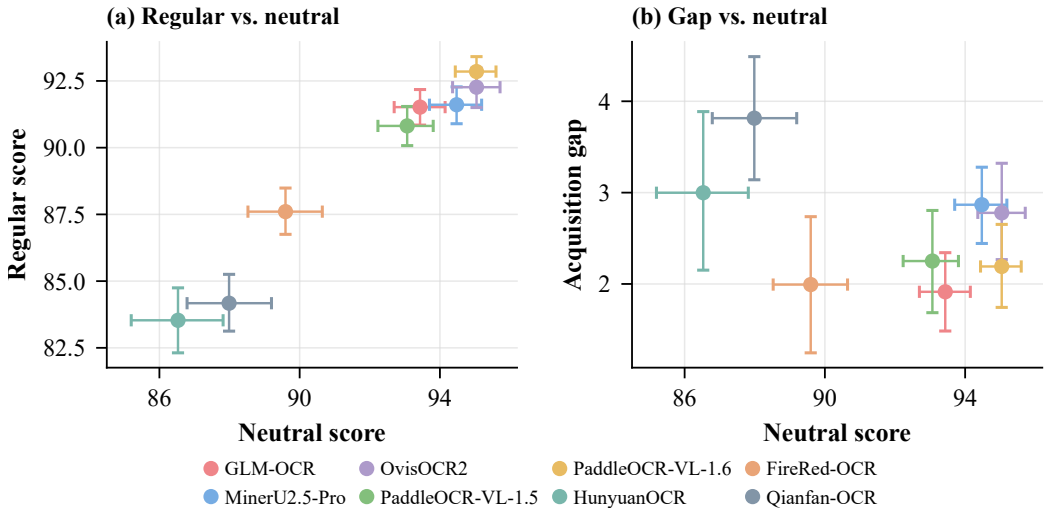


Figure 4: Neutral accuracy versus physical robustness. Left: regular versus neutral absolute accuracy (descriptive Spearman $\rho = 0.976$). Right: acquisition gap versus neutral accuracy ($\rho = -0.357$). Error bars are paired source-page bootstrap 95% intervals.

points for Qianfan-OCR (95% interval [4.58, 11.84]) and 6.01 for MinerU2.5-Pro ([3.35, 9.50]). Joint rectification and background replacement under vertical folds yield gains of 3.78 for Qianfan-OCR ([1.45, 6.89]) and 6.20 for HunyuanOCR ([2.52, 10.54]).

Transfer to Real Acquisition To assess how closely our simulation reproduces the effects of real acquisition on parser performance, we compare the same parsers on 1,355 matched Real5-OmniDocBench pages acquired under real-world conditions (Zhou et al., 2026) and simulated images of the same document contents. Real drops use Scanning as the reference; simulated drops use neutral images. We map real Skew to camera pitch, roll, and yaw; Warping to paper-geometry factors; and Illumination to illumination factors. Screen-Photography is compared with spotlight, ISO noise, and chromatic aberration as an indirect-coverage check. For each condition, we compare eight-parser rankings by aggregate drop using Kendall τ and paired source-page bootstrap intervals (Appendix D.2). Pose–Skew has $\tau = 0.79$ and geometry–Warping 0.36 (Table 3). Illumination has a small positive estimate, differences between our lighting presets and Real5's lighting may contribute to this weak agreement. Screen-Photography is near zero and our factors omit screen-specific artifacts such as moiré, which may explain the limited agreement.

Table 3: Real-to-sim Kendall τ on task-balanced drops (paired 95% intervals).

Real condition	τ [95% CI]
Skew	+0.79 [+0.50, +0.93]
Warping	+0.36 [+0.00, +0.71]
Illumination	+0.07 [-0.43, +0.36]
Screen-Photography	+0.00 [-0.21, +0.64]

7 Conclusion

We introduce SimDocBench, a physically simulated benchmark of full document parsing across 31 acquisition factors. Paired source, neutral, and perturbed images separate rendering losses from acquisition gaps. Eight parsers lose accuracy under changed capture conditions, and clean-page accuracy poorly predicts the size of that loss. The real-acquisition comparison shows where these patterns carry over to photographs. Beyond evaluation, the same physical simulation and rendering pipeline could augment OCR and document-parsing training data with varied in-the-wild acquisition conditions while retaining the source-page annotations. SimDocBench provides a controlled basis for diagnosing and improving document parsing under physical capture.

AI USE STATEMENT

The research ideas and experimental designs were developed by the authors. We used AI tools to assist with research implementation, including writing experimental scripts, optimizing ARCSim performance, and developing rendering configurations and components of the dataset synthesis pipeline. For manuscript preparation, we used AI tools to assist with translation, organization, language polishing, and LaTeX typesetting, including the revision of the Experiments, Further Analysis, and appendix sections. The authors reviewed and revised the AI-assisted text and code and take responsibility for the final content of this work, including its claims, results, and artifacts.

REFERENCES

Blender Foundation. Cycles renderer. Official source repository, n.d. URL https://github.com/blender/cycles. Accessed September 22, 2026.

Jean-Christophe Burie, Joseph Chazalon, Mickaël Coustaty, Sébastien Eskenazi, Muhammad Muzzamil Luqman, Maroua Mehri, Nibal Nayef, Jean-Marc Ogier, Sophea Prum, and Marçal Rusiñol. ICDAR2015 competition on smartphone document capture and OCR (smartdoc). In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pp. 1161–1165. IEEE Computer Society, 2015. doi: 10.1109/ICDAR.2015.7333943. URL https://doi.org/10.1109/ICDAR.2015.7333943.

Peng Cai, Zhaofan Zou, Shifa Liu, Yikun Wang, Jiawei Tang, Kaicheng Yang, Meng Tong, MingKun Jiang, Zhongjiang He, and Hao Sun. Teleocr: Navigating document parsing across digital and camera-captured documents, 2026. URL https://arxiv.org/abs/2608.12898.

Tsong Yueh Chen, Shing-Chi Cheung, and Siu-Ming Yiu. Metamorphic testing: A new approach for generating next test cases. *CoRR*, abs/2002.12543, 2020. URL https://arxiv.org/abs/2002.12543.

Yuxing Cheng, Yuan Wu, and Yi Chang. How robust is ocr-reasoning? evaluating ocr-reasoning robustness of vision-language models under visual perturbations. *CoRR*, abs/2606.26041, 2026. doi: 10.48550/ARXIV.2606.26041. URL https://doi.org/10.48550/arXiv.2606.26041.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr-vl-1.5: Towards a multi-task 0.9b VLM for robust in-the-wild document parsing. *CoRR*, abs/2601.21957, 2026. doi: 10.48550/ARXIV.2601.21957. URL https://doi.org/10.48550/arXiv.2601.21957.

Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 131–140. IEEE, 2019. doi: 10.1109/ICCV.2019.00022. URL https://doi.org/10.1109/ICCV.2019.00022.

Sagnik Das, Hassan A. Sial, Ke Ma, Ramón Baldrich, María Vanrell, and Dimitris Samaras. Intrinsic decomposition of document images in-the-wild. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0906.pdf.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, Yicheng Wang, Xue Xiong, Ayong Zheng, Xiaoying Zuo, Ziwei Ou, Jingnan Gu, Quanhao Guo, Jianmin Wu, Dawei Yin, and Dou Shen. Qianfan-ocr: A unified end-to-end model for document intelligence. *CoRR*, abs/2603.13398, 2026. doi: 10.48550/ARXIV.2603.13398. URL https://doi.org/10.48550/arXiv.2603.13398.

Yongkun Du, Pinxuan Chen, Xuye Ying, and Zhineng Chen. Docptbench: Benchmarking end-to-end photographed document parsing and translation. *CoRR*, abs/2511.18434, 2025. doi: 10.48550/ARXIV.2511.18434. URL https://doi.org/10.48550/arXiv.2511.18434.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, Dexin Jin, Yuxuan Zhang, Guohong Wen, Yanfeng Wang, Yutao Zhang, Xiaohan Zhang, Wenyi Hong, Yukuo Cen, Da Yin, Bin Chen, Wenmeng Yu, Xiaotao Gu, and Jie Tang. GLM-OCR technical report. *CoRR*, abs/2603.10910, 2026. doi: 10.48550/ARXIV.2603.10910. URL https://doi.org/10.48550/arXiv.2603.10910.

Anurag Dwarakanath, Manish Ahuja, Samarth Sikand, Raghotham M. Rao, R. P. Jagadeesh Chandra Bose, Neville Dubash, and Sanjay Podder. Identifying implementation bugs in machine learning based image classifiers using metamorphic testing. In Frank Tip and Eric Bodden (eds.), *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2018, Amsterdam, The Netherlands, July 16-21, 2018*, pp. 118–128. ACM, 2018. doi: 10.1145/3213846.3213858. URL https://doi.org/10.1145/3213846.3213858.

Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (eds.), *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 273–281. ACM, 2021. doi: 10.1145/3474085.3475388. URL https://doi.org/10.1145/3474085.3475388.

Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697 of *Lecture Notes in Computer Science*, pp. 475–492. Springer, 2022. doi: 10.1007/978-3-031-19836-6_27. URL https://doi.org/10.1007/978-3-031-19836-6_27.

Hao Feng, Wengang Zhou, Jiajun Deng, Qi Tian, and Houqiang Li. Docscanner: Robust document image rectification with progressive learning. *Int. J. Comput. Vis.*, 133(8):5343–5362, 2025. doi: 10.1007/S11263-025-02431-5. URL https://doi.org/10.1007/s11263-025-02431-5.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/8c2e6bb15be1894b8fb4e0f9bcad1739-Abstract-Datasets_and_Benchmarks_Track.html.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Gui-Song Xia. Revisiting document image dewarping by grid regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 4533–4542. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00450. URL https://doi.org/10.1109/CVPR52688.2022.00450.

James T. Kajiya. The rendering equation. In David C. Evans and Russell J. Athay (eds.), *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1986, Dallas, Texas, USA, August 18-22, 1986*, pp. 143–150. ACM, 1986. doi: 10.1145/15922.15902. URL https://doi.org/10.1145/15922.15902.

Gengluo Li, Pengyuan Lyu, Chengquan Zhang, Huawen Shen, Liang Wu, Xingyu Wan, Gangyan Zeng, Han Hu, Can Ma, and Yu Zhou. Towards real-world document parsing via realistic scene synthesis and document-aware training. *CoRR*, abs/2603.23885, 2026a. doi: 10.48550/ARXIV.2603.23885. URL https://doi.org/10.48550/arXiv.2603.23885.

Pu Li, Weize Quan, Jianwei Guo, and Dong-Ming Yan. Layout-aware single-image document flattening. *ACM Trans. Graph.*, 43(1):9:1–9:17, 2024. doi: 10.1145/3627818. URL https://doi.org/10.1145/3627818.

Zhang Li, Zhibo Lin, Qiang Liu, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiajun Song, Jiarui Zhang, Xiang Bai, and Yuliang Liu. Mdpbench: A benchmark for multilingual document parsing in real-world scenarios. *CoRR*, abs/2603.28130, 2026b. doi: 10.48550/ARXIV.2603.28130. URL https://doi.org/10.48550/arXiv.2603.28130.

Zhiheng Li, Zongyang Ma, Jiaxian Chen, Jianing Zhang, Zhaolong Su, Yutong Zhang, Zhiyin Yu, Ruiqi Liu, Xiaolei Lv, Bo Li, Jun Gao, Ziqi Zhang, Chunfeng Yuan, Bing Li, and Weiming Hu. How far is document parsing from solved? puredocbench: A source-traceablebenchmark across clean, degraded, and real-world settings. *CoRR*, abs/2605.07492, 2026c. doi: 10.48550/ARXIV.2605.07492. URL https://doi.org/10.48550/arXiv.2605.07492.

Meng Ling, Jian Chen, Torsten Möller, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Robert S. Laramee, Han-Wei Shen, Jian Wu, and C. Lee Giles. Document domain randomization for deep learning document layout extraction. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida (eds.), *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pp. 497–513. Springer, 2021. doi: 10.1007/978-3-030-86549-8_32. URL https://doi.org/10.1007/978-3-030-86549-8_32.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of OCR in large multimodal models. *Sci. China Inf. Sci.*, 67(12), 2024. doi: 10.1007/S11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Shiyin Lu, Yinglun Li, Yu Xia, Yuhui Chen, An-Yang Ji, Jun-Peng Jiang, Qing-Guo Chen, Jianshan Zhao, En Lin, Haijun Li, Cheng Qin, Zhao Xu, and Weihua Luo. Ovisocr2 technical report. *CoRR*, abs/2607.13639, 2026. doi: 10.48550/ARXIV.2607.13639. URL https://doi.org/10.48550/arXiv.2607.13639.

Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: Document image unwarping via a stacked u-net. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4700–4709. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00494. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Ma_DocUNet_Document_Image_CVPR_2018_paper.html.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 3571–3583, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1d49780520898fe37f0cd6b41c5311bf-Abstract.html.

Rahul Narain, Armin Samii, and James F. O'Brien. Adaptive anisotropic remeshing for cloth simulation. *ACM Trans. Graph.*, 31(6):152:1–152:10, 2012. doi: 10.1145/2366145.2366171. URL https://doi.org/10.1145/2366145.2366171.

Rahul Narain, Tobias Pfaff, and James F. O'Brien. Folding and crumpling adaptive sheets. *ACM Trans. Graph.*, 32(4):51:1–51:8, 2013. doi: 10.1145/2461912.2462010. URL https://doi.org/10.1145/2461912.2462010.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24838–24848. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02313. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ouyang_OmniDocBench_Benchmarking_Diverse_PDF_Document_Parsing_with_Comprehensive_Annotations_CVPR_2025_paper.html.

Chunyi Peng, Zhipeng Xu, Yuqi Xiong, Zulong Chen, Junhao Ji, Qing Liu, Zhenghao Liu, Zubao Qin, Bing Zhao, Jianqiang Wan, Jun Tang, Zhibo Yang, Shuai Bai, Dayiheng Liu, and Maosong Sun. Cc-ocr v2: Fine-grained attribution of lmm failures in real-world visual document understanding, 2026. URL https://arxiv.org/abs/2605.03903.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 9014–9023. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00843. URL https://doi.org/10.1109/ICCV51701.2025.00843.

Jenny Schmalfuss, Victor Oei, Lukas Mehl, Madlen Bartsch, Shashank Agnihotri, Margret Keuper, and Andrés Bruhn. Robustspring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo (1.0). https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/DARUS-5047{&}version=1.0, May 2025. URL https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/DARUS-5047&version=1.0.

Lloyd S Shapley et al. A value for n-person games. 1953.

Lin Sun, Dexian Wang, Jingang Huang, Linglin Zhang, Change Jia, Zhengwei Cheng, and Xiangzheng Zhang. When good OCR is not enough: Benchmarking OCR robustness for retrieval-augmented generation. In Yunyao Li, Georg Rehm, and Mei Tu (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), ACL 2026, San Diego, California, USA, July 2-7, 2026*, pp. 884–894. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-INDUSTRY.60. URL https://doi.org/10.18653/v1/2026.acl-industry.60.

Ting Sun, Cheng Cui, Yuning Du, and Yi Liu. Pp-doclayout: A unified document layout detection model to accelerate large-scale data construction. *CoRR*, abs/2503.17213, 2025. doi: 10.48550/ARXIV.2503.17213. URL https://doi.org/10.48550/arXiv.2503.17213.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, Qi Yang, Qiming Peng, Bin Luo, Hower Yang, Xinsong Zhang, Jinnian Zhang, Houwen Peng, Hongming Yang, Senhao Xie, Longsha Zhou, Ge Pei, Binghong Wu, Rui Yan, Kan Wu, Jieneng Yang, Bochao Wang, Kai Liu, Jianchen Zhu, Jie Jiang, Linus, Han Hu, and Chengquan Zhang. Hunyuanocr technical report. *CoRR*, abs/2511.19575, 2025. doi: 10.48550/ARXIV.2511.19575. URL https://doi.org/10.48550/arXiv.2511.19575.

Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. Deeptest: automated testing of deep-neural-network-driven autonomous cars. In Michel Chaudron, Ivica Crnkovic, Marsha Chechik, and Mark Harman (eds.), *Proceedings of the 40th International Conference on Software Engineering, ICSE 2018, Gothenburg, Sweden, May 27 - June 03, 2018*, pp. 303–314. ACM, 2018. doi: 10.1145/3180155.3180220. URL https://doi.org/10.1145/3180155.3180220.

Floor Verhoeven, Tanguy Magne, and Olga Sorkine-Hornung. Uvdoc: Neural grid-based document unwarping. In June Kim, Ming C. Lin, and Bernd Bickel (eds.), *SIGGRAPH Asia 2023 Conference Papers, SA 2023, Sydney, NSW, Australia, December 12-15, 2023*, pp. 110:1–110:11. ACM, 2023. doi: 10.1145/3610548.3618174. URL https://doi.org/10.1145/3610548.3618174.

An-Lan Wang, Jingqun Tang, Lei Liao, Hao Feng, Qi Liu, Xiang Fei, Jinghui Lu, Han Wang, Hao Liu, Yuliang Liu, Xiang Bai, and Can Huang. Wilddoc: How far are we from achieving comprehensive and robust document understanding in the wild? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 22991–23001. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1172. URL https://doi.org/10.18653/v1/2025.emnlp-main.1172.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, Bangrui Xu, Junbo Niu, Mengzhang Cai, Jiantao Qiu, Qintong Zhang, Dongsheng Ma, Yuefeng Sun, Hejun Dong, Wenzheng Zhang, Jutao Xiao, Jiayong Shi, Pengyu Liao, Xiaomeng Zhao, Huaping Zhong, Liqun Wei, Jing Yu, Jie Yang, Wei Li, Shasha Wang, Qianqian Wu, Xuanhe Zhou, Weijia Li, Zhenxiang Li, Zhongying Tu, Jiang Wu, Lijun Wu, Chao Xu, Kai Chen, Wentao Zhang, Yu Qiao, Bowen Zhou, Dahua Lin, and Conghui He. Mineru2.5-pro: Pushing the limits of data-centric document parsing at scale. *CoRR*, abs/2604.04771, 2026. doi: 10.48550/ARXIV.2604.04771. URL https://doi.org/10.48550/arXiv.2604.04771.

Hao Wu, Haoran Lou, Xinyue Li, Zuodong Zhong, Zhaojun Sun, Phellon Chen, Xuanhe Zhou, Kai Zuo, Yibo Chen, Xu Tang, Yao Hu, Boxiang Zhou, Jian Wu, Yongji Wu, Wenxin Yu, Yingmiao Liu, Yuhao Huang, Manjie Xu, Gang Liu, Yidong Ma, Zhichao Sun, and Changhao Qiao. Firered-ocr technical report. *CoRR*, abs/2603.01840, 2026. doi: 10.48550/ARXIV.2603.01840. URL https://doi.org/10.48550/arXiv.2603.01840.

Hao Yu, Kang Liu, Linnan Zhao, Jiabo Zhan, Chong Sun, Chen Li, and Jing Lyu. Wevisdoc: From coverage to capability for robust end-to-end document parsing, 2026. URL https://arxiv.org/abs/2609.20423.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, Tingquan Gao, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *CoRR*, abs/2606.03264, 2026. doi: 10.48550/ARXIV.2606.03264. URL https://doi.org/10.48550/arXiv.2606.03264.

Changda Zhou, Ziyue Gao, Xueqing Wang, Tingquan Gao, Cheng Cui, Jing Tang, and Yi Liu. Real5-omnidocbench: A full-scale physical reconstruction benchmark for robust document parsing in the wild. *CoRR*, abs/2603.04205, 2026. doi: 10.48550/ARXIV.2603.04205. URL https://doi.org/10.48550/arXiv.2603.04205.

A Overview of the Appendix

This appendix provides supporting details for SimDocBench. It is organized as follows:

- Appendix B describes the source pages, factor settings, compatibility rules, and construction of the SimDocBench-Regular and SimDocBench-Hard multi-factor settings.
- Appendix C specifies official task-balanced scoring, matched-neutral comparisons, and oracle restoration metrics.
- Appendix D reports supplementary intervals, component scores, factor-level results, coverage analyses, restoration diagnostics, and the real-acquisition ranking protocol.

B Dataset Details and Datasheet

We first document the source-page composition and rendered collections, then give the candidate settings and sampling constraints needed to reconstruct single-factor images and the SimDocBench-Regular and SimDocBench-Hard multi-factor settings.

B.1 Construction Protocol and Sampling

Source support and collections. The 1,651 source pages consist of 1,355 OmniDocBench v1.5 pages and 100 equation-hard, 99 layout-hard, and 97 table-hard pages. The language categories are English (755), Chinese (778), and mixed/other (118); document types include books (276), slides (253), academic papers (215), exam papers (193), textbooks (159), newspapers (151), magazines (149), research reports (132), notes (118), and historical documents (5). Evaluated isolated equations occur on 313 pages and tables on 458, with overlap. Each source has a canonical neutral render and the factor-appropriate controls described in Appendix C.2. Single-factor images and the SimDocBench-Regular and SimDocBench-Hard settings provide multiple acquisition views of shared source pages. The completed Hard collection contains one audited view of each of the 1,651 sources.

Single-factor settings. Tables 4 and 5 enumerate the candidate settings of every evaluated factor. The entries are candidate values, not frequencies or a shared severity scale. Thirty factors have three rendered variants per page, even when their candidate pool has more than three entries (buckling, crease, curl, watermark, and table texture); paper material has all six. Optical and lighting choices are often categorical rather than ordered. The nominal image canvas is 2048×1536 with 32 Cycles samples and content-dependent resolution adjustment as specified in Section 5.1; the motion-blur magnitude is capped at 3 pixels because manual inspection found stronger blur unsuitable for text recognition.

Tables 4 and 5 specify the candidate pools for all 31 single-factor interventions: 15 camera and illumination factors and 16 geometry, surface, and background factors. They show why the sampled variants must be interpreted as specified choices rather than equal steps on one severity scale; for example, the light-direction presets are categorical, while roll is measured in degrees.

For the standard bend, fold, and twist presets, the mild/medium/strong configurations use bending multipliers 0.7/1.0/1.3; these are mesh-simulation parameters, not measured optical severity. Twist also uses vertical displacement parameters 0.012/0.018/0.025. Other geometric settings additionally depend on mesh, location, aspect ratio, and precomputed-frame choice; their qualitative labels should not be treated as a shared physical scale.

Paper meshes and validation. Each source image is mapped to a triangular paper mesh with two-dimensional material (UV) coordinates and a nominal width of 0.21 m, preserving its aspect ratio. ARCSim simulates thin-sheet deformation under bending, stretching, contact, and prescribed constraints. We precompute grids and their deformation meshes for the source-page aspect ratios and select the corresponding mesh pool during rendering, avoiding repeated simulation for pages with the same aspect ratio.

We screen annotated regions for geometric orientation, self-occlusion, field-of-view coverage, and text scale, and rendered images for brightness. Surface normals and triangle areas provide an area-weighted view-angle distribution and front-facing fraction. For occlusion, we rasterize paper triangles

Table 4: Camera and illumination factor candidate settings in the single-factor recipes. Degrees are signed directions where applicable. The ISO names specify nominal presets, while the applied noise strengths are shown in parentheses.

Factor	Candidate settings
Chromatic aberration	1, 2, 3 px
Exposure multiplier	0.5, 1.0, 1.5
Focal length	28, 45, 70 mm
ISO noise	ISO 800, 1600, 3200 (strength 0.25, 0.45, 0.70)
Lens distortion coefficient	-0.02, -0.05 (barrel), $+0.02$ (pincushion)
Motion-blur magnitude	0.75, 1.5, 3 px
Motion-blur direction	0, 45, 90 degrees (at 2 px magnitude)
Pitch	forward 5, backward 5, forward 10 degrees
Roll	3, 5, 8 degrees
Vignetting strength	0.18, 0.30, 0.42
Yaw	left 5, right 5, left 10 degrees
Color temperature	6500, 5000, 3200 K
Light direction	left, right, back
Cast shadow	light, medium, heavy
Spotlight	centered, left-angled, narrow beam

Table 5: Geometry, surface, and background candidate settings.

Factor	Candidate settings
Concave, convex, compound bends (three factors)	mild, medium, strong (each)
Buckling	mild/medium/strong horizontal; medium vertical
Crease	three horizontal, two vertical, two diagonal positions/strengths
Curl	mild/medium/strong top; mild left; medium right
Dog ear	mild, medium, strong (content-safe candidates)
Horizontal fold, vertical fold, twist (three factors)	mild, medium, strong (each)
Wrinkles	physical mesh ranks 1, 2, 3
Grid underlay	faint, blue, dense (strength 0.18, 0.35, 0.50)
Splatter stains	light, medium, heavy (strength 0.20, 0.45, 0.70)
Watermark	seven text/strength presets (0.15–0.35)
Paper material	office matte, coated glossy, parchment, recycled rough, high-gloss photo, newsprint
Table texture	25 photographic/PBR surface presets

into a depth buffer, triangulate annotation polygons, and compare sampled surface depths with the frontmost surface; we also check that annotated UV regions remain in frame. We estimate effective character height from annotation area and non-whitespace character count, increasing rendering resolution when the estimated size falls below the minimum pixel requirement. Finally, mean brightness, dynamic range, and clipped-highlight fraction identify globally overexposed or otherwise unusable renders. The additional Hard-set content-region thresholds are specified below.

Regular multi-factor generation. The five regular views per page are sampled from 13 pools: deformation, exposure, focal length, optical artifacts, motion blur, pose, roll, lighting, paper material, grid, splatter, watermark, and table texture. The sampler draws a target of two to five active pools with activation weights, then one weighted compatible choice per active pool, leaving the others at baseline. It rejects incompatible selections, prevents duplicate full choice combinations for a page, and merges overrides in pool-priority order before rendering and visibility checks. Inactive and active candidates within the same pool cannot co-occur: the 11 geometry factors share one deformation pool; yaw/pitch share pose; four optical artifacts share one pool; four lighting scenarios share one pool; blur magnitude and direction are combined in one motion-blur choice. Cross-pool compatibility is checked for each sampled choice combination.

Cross-pool exclusions are choice-specific. In the regular recipe, nonbaseline focal length, lens distortion, yaw/pitch, and roll choices are excluded with the dog-ear, curl, or wrinkle choices requiring

pose-specific projected content fits; other deformation choices remain available. Some shadow choices exclude particular exposure or optical-artifact settings. The sampler checks compatibility of actual choices: if a chosen pool has no compatible active choice, that candidate combination is rejected and resampled. The stored image metadata records the sampled choices. An audit of all 8,255 images finds zero explicit choice-constraint violations. Projected onto the current 13-pool definitions, the images have 181, 2,236, 2,207, 2,172, and 1,459 active pools for counts one through five, respectively; changes to the pool definitions since generation account for the 181 images with one active current pool.

Hard acquisition recipe. The initial Hard draw uses a random-pose eight-pool recipe. It fixes physical wrinkle rank 2 and recycled-rough paper; independently samples roll and pitch from $\{-30, -20, -10, 10, 20, 30\}^\circ$, disallowing simultaneous 30° magnitudes. It also samples splatter (light/medium/heavy), one of 25 table materials, and no grid with probability 0.70 or one of three grids with probability 0.10 each. A 3-pixel blur is eligible only if the pose-adjusted estimated character width and height reach 5.5 and 14 pixels, respectively; 25% of eligible pages receive blur (232 of the 1,651 initial candidates under seed 123). All eight pools are present even if the chosen grid is "no grid" or blur is disabled. Pose-aware projections and content-fit checks permit roll/pitch with wrinkles in this recipe.

Each page's candidate wrinkle mesh is selected with a conservative pose preflight and checked again using its exact final camera projection. The final content-region screen requires every evaluated annotation to be in frame, have at least 24 projected pixels, at least 99.9% UV coverage, at least 98% depth-visible and front-facing area, and an area-weighted 95th-percentile view angle no greater than 65°. Of 1,651 initial candidates, 1,430 pass this screen. We redraw the 221 remaining sources with independent seeds and the same recipe; 204 pass. The final 17 use the existing gentle physical-wrinkle frames and $\pm 5^\circ$ roll/pitch settings, retaining the same content-region thresholds and other recipe pools. We keep the first accepted view per source, preferring the initial image, then original-strength retries, then the gentler draw.

C Evaluation Details

C.1 Accuracy Scoring Details

We use the OmniDocBench evaluator (Ouyang et al., 2025) to normalize and match predictions against the remapped ground truth $Y_{p,c}$ for each rendered image. Following its overall-score convention, $A_m(c) = [T_m(c) + F_m^{\mathrm{CDM}}(c) + B_m^{\mathrm{TEDS}}(c)]/3$, where the text, formula, and table task scores are on a 0–100 scale; reading order and structure-only TEDS are reported separately.

Paired Task Aggregation. For each comparison, both conditions use the same source-page identifiers. Within each task, multiple variants of a source page are averaged before accumulating task-specific sums and denominators, so each source page contributes equally. A reference formula or table page without a matched prediction contributes zero to that task; pages without the reference task are excluded from its denominator. We recompute the task scores and their equally weighted overall for each of 2,000 paired source-page bootstrap replicates. Averaging per-page overall scores instead would not yield the same estimand.

C.2 Paired Comparison Details

Neutral Control Assignments and Audits. Light direction, cast shadow, spotlight, buckling, crease, curling, folded corner, paper material, and wrinkling use dedicated neutral families; the remaining 22 factors share a flat-paper neutral family. The assigned controls were manually inspected to confirm that they provide appropriate factor-specific neutral references. PertDrop measures net differences against these controls under the sampled settings. Parameter analyses use recorded settings, with categorical variants analyzed as distinct choices. Hard multi-factor comparisons use the same 1,651 source pages for every reference condition.

Baseline Offsets. We record the offset between the canonical neutral reference and each matched control as

$$\mathrm{BaselineDelta}_{m,f} = A_m(f_0) - A_m(N_f). \tag{4}$$

BaselineDelta plus PertDrop equals the canonical-to-perturbed gap on the aggregate task support. This decomposition distinguishes the offset of the matched baseline from the perturbation cost measured against that baseline.

Supplementary Components. We report text, formula CDM, raw table TEDS, reading order, and table-structure scores separately where appropriate. No per-page overall retention ratio is defined by the task-balanced accuracy protocol. Improvements remain as negative aggregate gaps.

C.3 Metamorphic Consistency Sampling

MR1 isolates the effect of neutral rendering: for each of the 1,651 source pages, it pairs the original image with its canonical neutral render, giving one pair per page. MR2 instead isolates a single acquisition change relative to its factor-matched neutral control. To keep this prediction-consistency diagnostic balanced across factors and focused on content-preserving changes, we select 100 source pages per factor by a fixed hash of the factor and page identifier, without consulting parser outputs. We retain only variants whose recorded choices can be classified as mild or medium, excluding strong choices and choices without a recognized severity label. This protocol yields eligible pairs for 22 of the 31 factors; it does not evaluate the other nine factors under MR2. Each of the 22 eligible factors contributes 100 selected page–factor units, or 2,200 in total. A page can occur under several factors, and a selected page–factor unit can contain several eligible variants, each paired separately with its matched neutral. Consequently, MR2 comprises 4,608 pairs from 1,247 distinct source pages, not 2,200 independent pages.

We assign pages, rather than individual variants, to calibration or test sets using the same fixed hash-based rule for both relations (nominally 20%/80%); no source page crosses that boundary. MR1 has 345 calibration pairs/pages and 1,306 test pairs/pages. MR2 has 905 calibration pairs from 252 pages and 3,703 test pairs from 995 pages. These counts are per parser: all eight parsers use the same selected pairs. Calibration pages set the 95th-percentile violation thresholds separately for each parser, relation, and applicable output component; the reported MCD and violation estimates use only test pages. When a test page has multiple MR2 variants or factors, its pair values are averaged within the page before source-page bootstrap resampling. The 995 MR2 test pages are a subset of the 1,306 MR1 test pages, but the relations use different pair types and different test support; their mean MCD values should not be read as a paired comparison of perturbation severity.

C.4 Oracle Restoration Details

Oracle Restoration Conditions. For each page–factor–variant unit (p, f, v) in the diagnostic subset, we evaluate five inputs on a fixed 2048×1536 canvas. Let X_0 be the original perturbed input. X_L uses the projected paper mask to replace the surrounding scene with a uniform neutral background while retaining page position and shape. X_G uses retained UV correspondences to map the observed paper to a canonical rectangle, inpainting the vacated region to preserve the surrounding scene appearance. X_{LG} combines geometric rectification and neutral-background replacement. X_N is the canonical neutral reference X_{p,f_0}, resized to the same canvas if necessary; it is not the factor-matched control N_f. Each input has correspondingly transformed annotations Y_r, for $r \in \{0, L, G, LG, N\}$. These operations use renderer geometry and masks without providing reference transcriptions or semantic-region crops to the parser.

Recovery and Interaction Metrics. For a fixed model and factor on the 150-page matched support, let $a_r = A_m(r)$ be the official three-task aggregate for repair condition r. In particular, a_N is measured on the restoration canvas and is not replaced with a score from a different resolution setting. Background-only, geometry-only, and joint gains are $g_L = a_L - a_0$, $g_G = a_G - a_0$, and $g_{LG} = a_{LG} - a_0$. We define the interaction ι as the departure from additive gains. We then allocate joint recovery using two-operation Shapley values (Shapley et al., 1953): each operation receives its

Table 6: Regular acquisition: paired 95% source-page bootstrap intervals for the two gaps. Point estimates are in Table 1.

System	Δ_R [95% CI]	Δ_M [95% CI]
PaddleOCR-VL-1.6	1.08 [0.70, 1.52]	2.19 [1.74, 2.65]
OvisOCR2	0.76 [0.22, 1.32]	2.78 [2.27, 3.32]
MinerU2.5-Pro	1.19 [0.67, 1.77]	2.87 [2.44, 3.28]
GLM-OCR	2.07 [1.56, 2.59]	1.91 [1.48, 2.34]
PaddleOCR-VL-1.5	1.80 [1.20, 2.46]	2.25 [1.69, 2.80]
FireRed-OCR	3.43 [2.40, 4.46]	1.99 [1.25, 2.74]
Qianfan-OCR	5.90 [4.90, 6.91]	3.82 [3.14, 4.49]
HunyuanOCR	2.71 [1.41, 4.03]	3.00 [2.15, 3.89]

Table 7: Hard acquisition: one audited view per source, including 17 gentler views; N_H is scored on the same support. Brackets give paired 95% source-page bootstrap intervals; $\Delta_H = N_H - H$.

System	N_H [95% CI]	H [95% CI]	Δ_H [95% CI]
PaddleOCR-VL-1.6	95.04 [94.43, 95.58]	78.61 [76.93, 80.24]	16.44 [14.82, 18.06]
OvisOCR2	94.28 [93.50, 94.95]	76.41 [74.74, 78.02]	17.87 [16.23, 19.50]
MinerU2.5-Pro	94.48 [93.71, 95.22]	54.19 [52.43, 56.05]	40.29 [38.35, 42.10]
GLM-OCR	93.43 [92.67, 94.15]	72.48 [70.80, 74.18]	20.95 [19.40, 22.61]
PaddleOCR-VL-1.5	93.07 [92.23, 93.81]	78.06 [76.33, 79.77]	15.01 [13.35, 16.75]
FireRed-OCR	88.48 [87.20, 89.69]	68.72 [66.77, 70.58]	19.76 [17.79, 21.71]
Qianfan-OCR	87.37 [86.12, 88.53]	50.41 [48.36, 52.51]	36.95 [34.81, 39.03]
HunyuanOCR	86.35 [85.02, 87.58]	66.64 [64.69, 68.61]	19.71 [17.78, 21.58]

average marginal gain when applied before or after the other operation.

$$\iota = a_{LG} - a_L - a_G + a_0, \tag{5}$$

$$\phi_L = \tfrac{1}{2}\big[(a_L - a_0) + (a_{LG} - a_G)\big],$$

$$\phi_G = \tfrac{1}{2}\big[(a_G - a_0) + (a_{LG} - a_L)\big]. \tag{6}$$

The allocations satisfy $\phi_L + \phi_G = g_{LG}$; positive ι indicates a joint gain larger than the sum of the separate gains. We additionally report the total neutral gap $a_N - a_0$ and the residual after joint restoration $a_N - a_{LG}$. Signed gains capture both improvements and declines. These quantities measure observable recovery under the specified input operations. The geometry intervention combines UV rectification with resampling, background inpainting, and page repositioning.

D Additional Experimental Results

D.1 Extended Results

This subsection uses the same frozen exports and official task-balanced score as Section 5. We first examine uncertainty in the aggregate gaps, then locate factor-specific and component-specific effects, and finally report consistency and restoration diagnostics.

To assess whether the source-to-neutral rendering gap and the neutral-to-acquisition gap persist across sampled pages, Table 6 reports both Regular gaps and their paired source-page bootstrap intervals for every parser. Table 7 applies the same paired scoring to the audited Hard view of each source page and its matched Hard neutral reference; it lists both absolute scores and their difference. Every Regular gap interval is positive, and the Hard acquisition gaps span 15.01–40.29 points, far above the Regular acquisition gaps of 1.91–3.82 points.

To identify which output tasks contribute to these aggregate losses, Figure 5 decomposes the source-to-neutral and neutral-to-Regular gaps for all eight parsers into text edit accuracy, formula CDM, and table TEDS. Each point is a task-score difference on matched source pages; horizontal intervals come from paired source-page resampling.

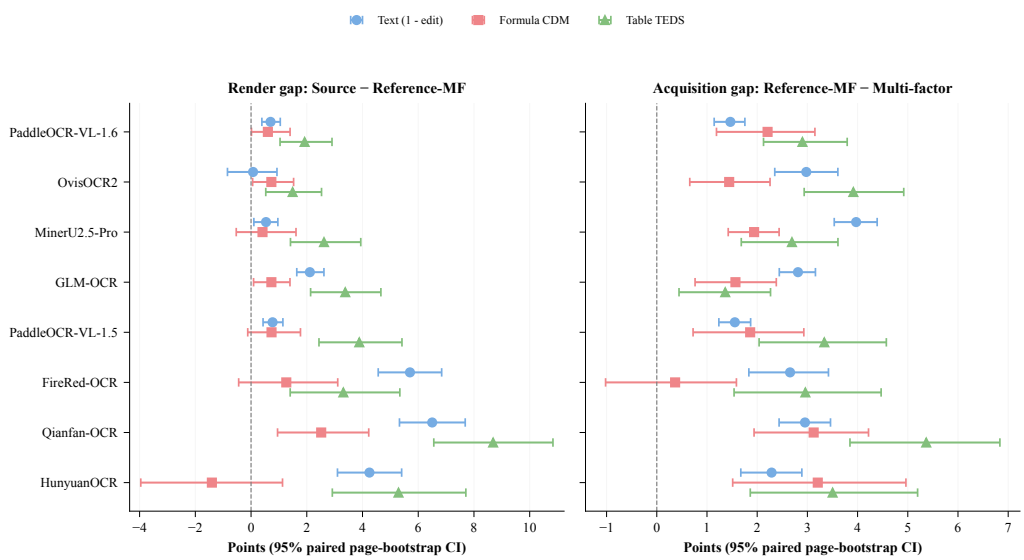


Figure 5: Component-level decomposition of the two gaps in Table 1. Each system shows text (1−edit), formula CDM, and table TEDS gaps with 95% paired source-page bootstrap intervals, for the rendering gap (left) and the acquisition gap (right).

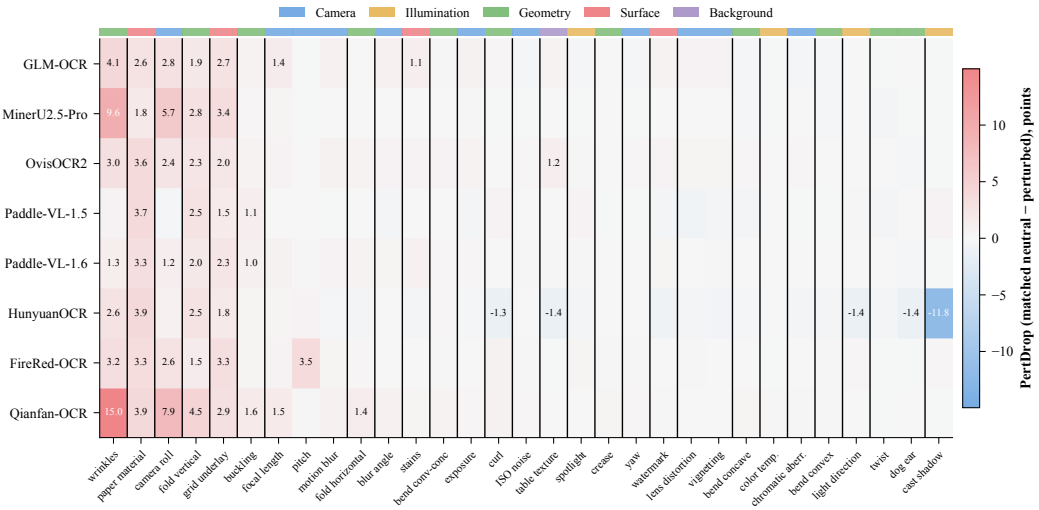


Figure 6: Full 31-factor by eight-parser PertDrop heatmap under the official three-task equally weighted score. Factors are ordered by their eight-parser mean drop; color strips identify physical factor groups. Positive values indicate lower accuracy than the factor-matched neutral control.

In Figure 5, table TEDS accounts for the largest acquisition-gap component in most systems, whereas some formula CDM gap intervals include zero.

To show whether the leading factors in the main-text plot represent a broad pattern or isolated parser responses, Figure 6 displays all 31 factors across eight parsers. Table 8 gives the corresponding task-balanced PertDrop values and 95% paired source-page bootstrap intervals against each factor's matched neutral control. The heatmap orders factors by their eight-parser mean and marks their physical groups; the table retains every parser–factor cell. Wrinkles produce large drops for several parsers, whereas the negative cast-shadow mean is driven mainly by HunyuanOCR's −11.84-point matched-neutral difference, which should be read relative to that factor's control.

Table 8: Full 31-factor PertDrop table. Each cell is the aggregate official three-task score difference against the matched-neutral control, with 95% paired source-page bootstrap intervals; factors are sorted by the eight-model mean drop.

Factor	GLM-OCR	MinerU2.5-Pro	OvisOCR2	PaddleOCR-VL-1.5	PaddleOCR-VL-1.6	HunyuanOCR	FireRed-OCR	Qianfan-OCR
wrinkles (Geometry)	4.09 [3.58, 4.61]	9.56 [8.79, 10.33]	3.02 [2.35, 3.70]	0.50 [-0.05, 1.05]	1.27 [0.80, 1.75]	2.56 [1.39, 3.64]	3.18 [2.20, 4.11]	14.97 [13.51, 16.58]
paper material (Surface)	2.61 [2.30, 2.92]	1.78 [1.45, 2.09]	3.64 [3.28, 4.03]	3.73 [3.29, 4.19]	3.27 [2.94, 3.61]	3.93 [2.95, 4.97]	3.35 [2.74, 3.89]	3.86 [3.46, 4.28]
camera roll (Camera)	2.82 [2.34, 3.28]	5.74 [5.15, 6.35]	2.44 [1.83, 3.01]	-0.30 [-0.88, 0.27]	1.16 [0.63, 1.73]	0.84 [-0.41, 1.98]	2.64 [1.85, 3.43]	7.94 [7.01, 8.87]
vertical fold (Geometry)	1.93 [1.43, 2.45]	2.80 [2.24, 3.38]	2.29 [1.43, 3.17]	2.52 [1.78, 3.22]	1.98 [1.38, 2.66]	2.49 [1.11, 3.85]	1.49 [0.65, 2.24]	4.54 [3.48, 5.57]
grid underlay (Surface)	2.65 [2.00, 3.36]	3.37 [2.78, 3.98]	2.02 [1.50, 2.59]	1.52 [0.70, 2.34]	2.28 [1.68, 2.93]	1.78 [0.64, 2.85]	3.26 [2.36, 4.11]	2.95 [2.18, 3.79]
buckling (Geometry)	0.65 [0.30, 1.02]	0.30 [-0.11, 0.70]	0.66 [0.10, 1.20]	1.14 [0.62, 1.64]	1.05 [0.70, 1.42]	0.38 [-0.89, 1.54]	0.45 [-0.23, 1.05]	1.57 [0.89, 2.25]
focal length (Camera)	1.41 [1.06, 1.76]	0.39 [0.09, 0.68]	0.58 [0.17, 0.99]	0.05 [-0.55, 0.62]	0.70 [0.35, 1.08]	0.37 [-0.66, 1.33]	0.57 [0.07, 1.07]	1.54 [1.11, 1.98]
pitch (Camera)	-0.17 [-0.46, 0.10]	-0.16 [-0.46, 0.11]	0.10 [-0.28, 0.48]	-0.03 [-0.48, 0.41]	0.15 [-0.18, 0.49]	0.51 [-0.46, 1.46]	3.55 [2.23, 4.95]	0.21 [-0.29, 0.72]
motion blur (Camera)	0.87 [0.54, 1.19]	0.06 [-0.16, 0.27]	0.75 [0.35, 1.17]	-0.20 [-0.76, 0.33]	0.56 [0.27, 0.88]	-0.25 [-1.23, 0.73]	0.40 [-0.14, 0.99]	0.91 [0.44, 1.43]
fold horizontal (Geometry)	-0.04 [-0.42, 0.32]	-0.02 [-0.43, 0.35]	0.68 [0.21, 1.16]	0.11 [-0.42, 0.61]	0.70 [0.29, 1.08]	-0.44 [-1.53, 0.61]	0.30 [-0.31, 0.89]	1.40 [0.81, 1.99]
motion blur angle (Camera)	0.90 [0.58, 1.23]	0.08 [-0.21, 0.34]	0.47 [0.07, 0.84]	-0.36 [-0.97, 0.23]	0.45 [0.11, 0.80]	-0.05 [-1.11, 0.98]	0.10 [-0.30, 0.54]	0.93 [0.45, 1.47]
stain splatter (Surface)	1.12 [0.75, 1.50]	0.02 [-0.16, 0.19]	0.58 [0.22, 0.99]	-0.04 [-0.69, 0.58]	0.92 [0.52, 1.33]	-0.55 [-1.46, 0.38]	-0.02 [-0.47, 0.40]	0.43 [0.07, 0.83]
bend convex concave (Geometry)	0.18 [-0.20, 0.53]	-0.05 [-0.36, 0.25]	0.52 [0.04, 0.99]	0.09 [-0.49, 0.63]	0.31 [-0.10, 0.68]	0.31 [-0.79, 1.38]	0.22 [-0.48, 0.89]	0.62 [0.11, 1.12]
exposure (Camera)	0.82 [0.53, 1.11]	0.14 [-0.02, 0.31]	0.65 [0.25, 1.08]	-0.56 [-1.16, -0.03]	0.30 [-0.02, 0.61]	-0.32 [-1.28, 0.56]	0.09 [-0.37, 0.52]	0.25 [-0.05, 0.59]
curl (Geometry)	0.24 [-0.08, 0.58]	-0.17 [-0.50, 0.17]	-0.08 [-0.65, 0.37]	0.48 [0.04, 0.96]	0.01 [-0.42, 0.42]	-1.29 [-2.48, -0.17]	0.74 [0.14, 1.37]	0.73 [0.13, 1.30]
iso noise (Camera)	-0.25 [-0.55, 0.02]	-0.15 [-0.35, 0.05]	0.31 [0.01, 0.62]	0.00 [-0.45, 0.43]	0.23 [-0.11, 0.56]	-0.08 [-1.23, 1.01]	0.20 [-0.50, 0.89]	0.32 [-0.07, 0.75]
table texture (Background)	0.60 [0.30, 0.93]	0.00 [-0.25, 0.22]	1.16 [0.68, 1.66]	-0.09 [-0.57, 0.34]	0.18 [-0.18, 0.52]	-1.38 [-2.50, -0.33]	-0.12 [-0.61, 0.37]	0.11 [-0.37, 0.60]
spotlight (Illumination)	-0.26 [-0.59, 0.04]	-0.15 [-0.41, 0.11]	0.14 [-0.18, 0.39]	0.48 [0.03, 0.89]	0.17 [-0.18, 0.49]	-0.43 [-1.61, 0.68]	0.36 [-0.28, 1.06]	-0.03 [-0.46, 0.40]
crease (Geometry)	0.01 [-0.13, 0.16]	-0.07 [-0.25, 0.10]	-0.01 [-0.20, 0.19]	-0.18 [-0.44, 0.08]	0.05 [-0.09, 0.21]	-0.11 [-0.77, 0.54]	0.17 [-0.18, 0.56]	0.38 [0.07, 0.72]
yaw (Camera)	-0.27 [-0.58, 0.03]	-0.21 [-0.55, 0.10]	0.24 [-0.11, 0.58]	-0.14 [-0.59, 0.29]	0.11 [-0.25, 0.46]	0.04 [-1.07, 1.09]	0.14 [-0.58, 0.88]	0.26 [-0.28, 0.86]
watermark (Surface)	0.59 [0.30, 0.89]	-0.05 [-0.27, 0.15]	0.55 [0.12, 0.95]	-0.57 [-1.22, 0.03]	0.40 [0.09, 0.72]	-0.90 [-1.92, 0.12]	-0.08 [-0.63, 0.43]	0.22 [-0.13, 0.62]
lens distortion (Camera)	0.58 [0.24, 0.92]	-0.09 [-0.40, 0.20]	0.46 [0.05, 0.86]	-0.75 [-1.41, -0.15]	0.11 [-0.24, 0.44]	-0.57 [-1.58, 0.44]	0.23 [-0.23, 0.75]	0.09 [-0.37, 0.55]
vignetting (Camera)	0.53 [0.26, 0.82]	-0.01 [-0.19, 0.16]	0.38 [-0.08, 0.84]	-0.48 [-1.06, 0.07]	0.14 [-0.16, 0.43]	-0.59 [-1.50, 0.27]	-0.08 [-0.49, 0.33]	0.04 [-0.20, 0.26]
bend concave (Geometry)	-0.26 [-0.60, 0.08]	-0.24 [-0.61, 0.10]	0.28 [-0.11, 0.64]	-0.36 [-0.89, 0.14]	0.09 [-0.27, 0.45]	-0.24 [-1.37, 0.81]	0.09 [-0.57, 0.71]	0.46 [-0.06, 0.99]
color temperature (Illumination)	-0.20 [-0.38, -0.05]	-0.13 [-0.31, 0.02]	0.00 [-0.38, 0.40]	-0.09 [-0.38, 0.20]	-0.02 [-0.30, 0.24]	0.06 [-0.79, 0.96]	-0.15 [-0.64, 0.28]	0.15 [-0.08, 0.39]
chromatic aberration (Camera)	-0.27 [-0.56, -0.03]	-0.03 [-0.19, 0.11]	0.33 [-0.07, 0.76]	-0.23 [-0.53, 0.08]	-0.01 [-0.33, 0.28]	-0.41 [-1.31, 0.50]	-0.03 [-0.62, 0.50]	0.23 [-0.20, 0.67]
bend convex (Geometry)	-0.29 [-0.67, 0.07]	-0.31 [-0.58, -0.04]	-0.09 [-0.51, 0.28]	-0.23 [-0.75, 0.25]	0.03 [-0.36, 0.39]	-0.37 [-1.46, 0.66]	0.15 [-0.37, 0.67]	0.14 [-0.39, 0.69]
direction (Illumination)	-0.07 [-0.31, 0.16]	-0.01 [-0.22, 0.19]	0.06 [-0.30, 0.33]	0.18 [-0.20, 0.53]	-0.06 [-0.44, 0.27]	-1.37 [-2.34, -0.42]	0.29 [-0.17, 0.70]	-0.01 [-0.24, 0.22]
twist (Geometry)	-0.20 [-0.58, 0.17]	-0.42 [-0.78, -0.10]	-0.00 [-0.42, 0.36]	-0.46 [-1.07, 0.11]	0.02 [-0.38, 0.41]	-0.56 [-1.76, 0.53]	-0.14 [-0.79, 0.46]	-0.17 [-0.74, 0.37]
dog ear (Geometry)	-0.05 [-0.31, 0.23]	-0.21 [-0.41, -0.02]	-0.24 [-0.70, 0.16]	0.02 [-0.34, 0.37]	-0.31 [-0.71, 0.07]	-1.36 [-2.22, -0.45]	-0.22 [-0.67, 0.19]	0.12 [-0.19, 0.45]
cast shadow (Illumination)	-0.23 [-0.61, 0.09]	-0.17 [-0.37, 0.03]	-0.17 [-0.56, 0.15]	0.52 [0.06, 0.96]	-0.22 [-0.60, 0.13]	-11.84 [-13.63, -10.01]	0.30 [-0.32, 0.99]	-0.13 [-0.51, 0.23]

Table 9: Group-level mean PertDrop by system (equal factor weights within groups) with 95% source-cluster bootstrap intervals.

Group	GLM-OCR	MinerU2.5-Pro	OvisOCR2	PaddleOCR-VL-1.5	PaddleOCR-VL-1.6	HunyuanOCR	FireRed-OCR	Qianfan-OCR
Camera	0.63 [0.37, 0.87]	0.52 [0.33, 0.70]	0.61 [0.27, 0.93]	-0.27 [-0.70, 0.13]	0.35 [0.07, 0.64]	-0.05 [-0.91, 0.79]	0.71 [0.25, 1.16]	1.16 [0.80, 1.52]
Illumination	-0.19 [-0.37, -0.03]	-0.12 [-0.27, 0.02]	0.01 [-0.18, 0.18]	0.27 [-0.02, 0.53]	-0.03 [-0.29, 0.19]	-3.40 [-4.12, -2.69]	0.20 [-0.10, 0.55]	-0.00 [-0.20, 0.19]
Geometry	0.57 [0.34, 0.78]	1.02 [0.77, 1.26]	0.64 [0.32, 0.93]	0.33 [-0.00, 0.65]	0.47 [0.23, 0.71]	0.12 [-0.64, 0.83]	0.58 [0.25, 0.91]	2.25 [1.85, 2.65]
Surface	1.74 [1.44, 2.06]	1.28 [1.05, 1.50]	1.70 [1.40, 2.00]	1.16 [0.66, 1.66]	1.72 [1.44, 2.00]	1.06 [0.27, 1.87]	1.63 [1.18, 2.05]	1.87 [1.54, 2.20]
Background	0.60 [0.30, 0.93]	0.00 [-0.25, 0.22]	1.16 [0.68, 1.66]	-0.09 [-0.57, 0.34]	0.18 [-0.18, 0.52]	-1.38 [-2.50, -0.33]	-0.12 [-0.61, 0.37]	0.11 [-0.37, 0.60]

To compare broader acquisition families without letting groups with more factors dominate, Table 9 averages factor-level PertDrops with equal factor weights within Camera, Illumination, Geometry, Surface, and Background for each parser. Its intervals resample source pages. Surface has positive mean drops for all eight parsers, while the Illumination mean is strongly negative for HunyuanOCR (-3.40), showing that group averages can still reflect a model-specific response.

Coverage Across Acquisition Factor Sets We examine whether the paper geometry and material factors modeled by our physical simulation and rendering expose parsing failures that are not captured by image-operator factors. For this comparison, *image-operator factors* comprise the 11 Camera factors, four Illumination factors, grid underlay, stain splatter, watermark, and background texture (19 factors). *Paper geometry and material factors* comprise all 11 Paper Geometry factors and paper material (12 factors); Figure 1 lists the factors within each category. The split specifies a planar-image-operator scope for comparison: planar transforms, brightness fields, blur/noise, and compositing fall in the image-operator set, including cast shadow as a shadow mask. Nonplanar paper shape and material appearance fall in the paper geometry and material set. Crease and paper material are boundary cases; the accompanying sensitivity analysis reassigns both to the image-operator set (21/10). This grouping does not establish that the paper geometry and material effects require a three-dimensional renderer.

We define a hit for a parser–factor pair when its official task-balanced matched-neutral PertDrop reaches five points over 1,651 pages. A factor set covers a parser if at least one of its factors yields a hit. The event is defined for a fixed parser–factor pair, not for a parser–page pair. Table 10 reports eight-parser macro coverage for the image-operator set, additional coverage from paper geometry and material, and their union, both with all factors and under equal-count factor subsampling. With all factors included, both sets cover 25.00% of parsers in the eight-parser macro average; the paper geometry and material set adds no parser at this threshold. Under equal-count factor subsampling, the paper-only coverage probability is 9.21% [9.21, 17.11]%. These probabilities concern sampled factor subsets for fixed parsers; their intervals resample source pages to re-estimate factor-level drops. They cannot be compared numerically with earlier page-event coverage percentages. Thresholds 1, 2, 5, 10, and 20 and a boundary-factor reclassification appear in the accompanying machine-readable results.

Table 10: Official task-balanced factor-level coverage at a five-point aggregate drop threshold. Image-operator coverage and additional paper geometry and material coverage are eight-parser macro averages with paired source-page bootstrap intervals; equal factor budgets sample subsets without replacement.

Factor budget	Image-operator factors	Paper geometry and material only	Union
full	25.00 [25.00, 25.00]	0.00 [0.00, 12.50]	25.00 [25.00, 37.50]
equal factor budget	15.79 [15.79, 15.79]	9.21 [9.21, 17.11]	25.00 [25.00, 32.89]

Table 11: Per-system official three-task source, matched-neutral, and regular multi-factor scores and acquisition gaps with 95% paired source-page bootstrap intervals.

Model	Source	Neutral	Regular	Acq. gap
GLM-OCR	95.51 [94.93, 96.02]	93.43 [92.70, 94.15]	91.52 [90.85, 92.18]	1.91 [1.48, 2.34]
MinerU2.5-Pro	95.66 [95.05, 96.21]	94.48 [93.70, 95.19]	91.61 [90.90, 92.28]	2.87 [2.44, 3.28]
OvisOCR2	95.81 [95.26, 96.32]	95.04 [94.36, 95.71]	92.27 [91.50, 92.98]	2.78 [2.27, 3.32]
PaddleOCR-VL-1.5	94.86 [94.20, 95.44]	93.07 [92.23, 93.81]	90.82 [90.08, 91.54]	2.25 [1.69, 2.80]
PaddleOCR-VL-1.6	96.12 [95.72, 96.50]	95.04 [94.44, 95.60]	92.85 [92.24, 93.41]	2.19 [1.74, 2.65]
HunyuanOCR	89.24 [87.97, 90.51]	86.53 [85.20, 87.82]	83.53 [82.31, 84.75]	3.00 [2.15, 3.89]
FireRed-OCR	93.02 [92.16, 93.91]	89.60 [88.53, 90.65]	87.60 [86.75, 88.49]	1.99 [1.25, 2.74]
Qianfan-OCR	93.89 [93.06, 94.67]	87.99 [86.79, 89.20]	84.17 [83.13, 85.25]	3.82 [3.14, 4.49]

To compare absolute parsing quality with sensitivity to Regular acquisition, Table 11 places each parser's source, canonical neutral, and Regular scores beside its neutral-to-Regular gap. All columns use the same three-task aggregation and paired source-page bootstrap support. PaddleOCR-VL-1.6 has the highest Regular score (92.85), whereas GLM-OCR has the smallest acquisition gap (1.91), confirming that absolute accuracy and the size of the acquisition loss give different rankings.

To examine whether prediction changes appear across text, formula, table, or reading-order outputs, Table 12 reports both continuous overall MCD and calibrated violation rates for each parser and metamorphic relation. Rendering consistency compares source images with canonical neutral renders (MR1), whereas acquisition consistency compares factor-matched neutral renders with single-factor perturbations (MR2; Section 4.3). On held-out pages, each model has 1,306 MR1 pairs from 1,306 source pages and 3,703 MR2 pairs from 995 source pages. The distinct sampling designs and calibration counts are detailed in Appendix C.3; violation thresholds are calibrated separately by model, relation, and output component. The table includes pair/page counts, overall MCD and any-component violation rates with intervals, followed by text, formula, table, and order violation rates. Under MR2, HunyuanOCR has the largest mean MCD (0.204) but an any-component violation rate of 0.061, whereas OvisOCR2 has the smallest mean MCD (0.054) and a rate of 0.122; the two summaries capture different aspects of consistency.

To test how much of each parser's loss is recoverable through controlled input corrections, Table 13 reports the official task-balanced score gain from background-only replacement (L), geometry-only rectification (G), and their combination (LG) relative to the original perturbed input. It also gives the gap to the canonical neutral reference on the same restoration canvas and the residual gap after joint correction, with paired source-page bootstrap intervals for every parser–factor pair. The corrections use renderer masks and UV correspondences on the matched 150-page diagnostic subset (Appendix C.4). Camera-roll geometry correction yields 7.97 points for Qianfan-OCR, while the joint vertical-fold correction yields 6.20 for HunyuanOCR; other entries show smaller or negative gains, so recoverability depends on the factor and parser.

D.2 Real-Acquisition Ranking Protocol

Real5-OmniDocBench provides 1,355 pages under Scanning, Skew, Warping, Illumination, and Screen-Photography (Zhou et al., 2026). We use the same eight parsers and official task-balanced score on both datasets, restricting simulated pages to the corresponding 1,355 source contents. For

Table 12: Per-system rendering and acquisition consistency: mean overall MCD with 95% confidence intervals, any-component violation rates, and per-component violation rates with intervals (Text, Formula, Table, Order). Component entries here are thresholded rates, not the mean distances in Table 2.

MR	Model	Pairs / pages	MCD [95% CI]	Any violation [95% CI]	Text	Formula	Table	Order
MR1	FireRed-OCR	1,306 / 1,306	0.103 [0.095, 0.111]	0.097 [0.082, 0.113]	0.054 [0.042, 0.066]	0.077 [0.055, 0.100]	0.029 [0.013, 0.048]	0.043 [0.032, 0.054]
MR1	GLM-OCR	1,306 / 1,306	0.170 [0.161, 0.181]	0.074 [0.061, 0.089]	0.045 [0.034, 0.057]	0.000 [0.000, 0.000]	0.039 [0.021, 0.059]	0.051 [0.040, 0.063]
MR1	HunyuanOCR	1,306 / 1,306	0.298 [0.285, 0.311]	0.066 [0.053, 0.080]	0.039 [0.028, 0.050]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.040 [0.030, 0.051]
MR1	MinerU2.5-Pro	1,306 / 1,306	0.128 [0.120, 0.137]	0.058 [0.046, 0.071]	0.042 [0.032, 0.053]	0.000 [0.000, 0.000]	0.042 [0.023, 0.062]	0.034 [0.025, 0.045]
MR1	OvisOCR2	1,306 / 1,306	0.105 [0.096, 0.115]	0.083 [0.068, 0.099]	0.070 [0.057, 0.084]	0.000 [0.000, 0.000]	0.036 [0.018, 0.054]	0.060 [0.047, 0.073]
MR1	Paddle-VL-1.5	1,306 / 1,306	0.133 [0.124, 0.141]	0.083 [0.069, 0.098]	0.052 [0.041, 0.064]	0.000 [0.000, 0.000]	0.036 [0.018, 0.054]	0.045 [0.034, 0.056]
MR1	Paddle-VL-1.6	1,306 / 1,306	0.118 [0.110, 0.125]	0.074 [0.060, 0.089]	0.041 [0.031, 0.052]	0.000 [0.000, 0.000]	0.036 [0.018, 0.054]	0.045 [0.034, 0.057]
MR1	Qianfan-OCR	1,306 / 1,306	0.163 [0.152, 0.176]	0.106 [0.090, 0.123]	0.058 [0.047, 0.071]	0.084 [0.061, 0.110]	0.055 [0.031, 0.081]	0.062 [0.051, 0.076]
MR2	FireRed-OCR	3,703 / 995	0.072 [0.065, 0.080]	0.096 [0.081, 0.111]	0.042 [0.032, 0.053]	0.083 [0.060, 0.108]	0.085 [0.059, 0.115]	0.049 [0.038, 0.060]
MR2	GLM-OCR	3,703 / 995	0.103 [0.094, 0.112]	0.075 [0.060, 0.090]	0.055 [0.043, 0.069]	0.000 [0.000, 0.000]	0.044 [0.024, 0.069]	0.056 [0.044, 0.070]
MR2	HunyuanOCR	3,703 / 995	0.204 [0.189, 0.219]	0.061 [0.049, 0.075]	0.037 [0.027, 0.047]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.050 [0.039, 0.061]
MR2	MinerU2.5-Pro	3,703 / 995	0.075 [0.069, 0.082]	0.071 [0.059, 0.084]	0.048 [0.038, 0.059]	0.000 [0.000, 0.000]	0.093 [0.066, 0.123]	0.038 [0.030, 0.048]
MR2	OvisOCR2	3,703 / 995	0.054 [0.047, 0.062]	0.122 [0.105, 0.138]	0.069 [0.055, 0.082]	0.083 [0.061, 0.106]	0.071 [0.048, 0.095]	0.078 [0.064, 0.092]
MR2	Paddle-VL-1.5	3,703 / 995	0.077 [0.071, 0.083]	0.078 [0.065, 0.092]	0.046 [0.036, 0.057]	0.000 [0.000, 0.000]	0.054 [0.032, 0.078]	0.040 [0.030, 0.050]
MR2	Paddle-VL-1.6	3,703 / 995	0.074 [0.068, 0.080]	0.069 [0.057, 0.082]	0.042 [0.033, 0.053]	0.000 [0.000, 0.000]	0.037 [0.019, 0.059]	0.035 [0.026, 0.045]
MR2	Qianfan-OCR	3,703 / 995	0.127 [0.114, 0.140]	0.107 [0.089, 0.123]	0.061 [0.049, 0.073]	0.096 [0.072, 0.125]	0.057 [0.034, 0.082]	0.068 [0.054, 0.082]

Table 13: Per-system, per-factor official task-balanced restoration results: background-only, geometry-only, and joint gains with 95% paired source-page bootstrap intervals, together with neutral-reference gaps and post-repair residuals.

Factor	Model	L	G	LG	Gap	Residual
camera roll	GLM-OCR	-0.23 [-0.76, 0.10]	2.90 [1.84, 3.97]	2.84 [1.77, 3.90]	2.45 [1.45, 3.34]	-0.39 [-1.11, 0.21]
camera roll	MinerU2.5-Pro	0.25 [-0.23, 0.79]	6.01 [3.35, 9.50]	5.87 [3.20, 9.33]	5.00 [1.59, 8.59]	-0.87 [-3.73, 1.19]
camera roll	OvisOCR2	-0.06 [-0.32, 0.13]	2.09 [-0.82, 4.57]	2.21 [-0.66, 4.71]	1.91 [-1.02, 4.23]	-0.30 [-1.02, 0.34]
camera roll	PaddleOCR-VL-1.5	0.00 [-0.54, 0.57]	0.39 [-0.98, 1.74]	0.27 [-1.11, 1.62]	1.04 [-0.25, 2.37]	0.77 [0.06, 1.58]
camera roll	PaddleOCR-VL-1.6	-0.02 [-0.44, 0.45]	0.75 [-0.07, 1.55]	0.68 [-0.14, 1.48]	0.60 [-0.32, 1.56]	-0.08 [-0.45, 0.26]
camera roll	HunyuanOCR	1.82 [-1.01, 5.00]	6.14 [2.10, 10.71]	5.04 [1.25, 9.77]	2.62 [-1.89, 7.56]	-2.42 [-6.58, 1.62]
camera roll	FireRed-OCR	-0.83 [-1.91, -0.08]	2.08 [-0.65, 4.54]	3.05 [1.43, 4.94]	3.10 [1.08, 5.51]	0.05 [-1.49, 1.83]
camera roll	Qianfan-OCR	0.73 [-2.17, 4.30]	7.97 [4.58, 11.84]	8.03 [4.69, 11.91]	8.13 [4.77, 11.93]	0.10 [-1.36, 1.52]
vertical fold	GLM-OCR	-0.38 [-1.06, 0.12]	2.09 [0.60, 3.57]	2.09 [0.61, 3.57]	1.90 [0.41, 3.30]	-0.19 [-1.14, 0.74]
vertical fold	MinerU2.5-Pro	-0.12 [-0.55, 0.25]	2.11 [0.91, 3.50]	1.91 [0.63, 3.27]	0.85 [-1.99, 2.74]	-1.06 [-3.94, 0.93]
vertical fold	OvisOCR2	-0.33 [-0.77, -0.01]	1.34 [-2.18, 4.81]	1.71 [-1.49, 4.88]	1.69 [-1.55, 4.60]	-0.01 [-0.82, 0.74]
vertical fold	PaddleOCR-VL-1.5	-0.07 [-0.96, 0.93]	2.21 [0.25, 4.37]	2.06 [0.11, 4.26]	2.86 [0.98, 4.88]	0.80 [0.05, 1.60]
vertical fold	PaddleOCR-VL-1.6	-0.62 [-2.01, 0.61]	1.38 [-1.06, 3.82]	1.33 [-1.17, 3.76]	1.57 [-0.85, 4.05]	0.24 [-0.37, 0.92]
vertical fold	HunyuanOCR	4.52 [1.22, 8.59]	8.05 [3.82, 12.84]	6.20 [2.52, 10.54]	4.83 [-0.13, 10.31]	-1.37 [-5.44, 2.74]
vertical fold	FireRed-OCR	-0.35 [-1.30, 0.65]	3.35 [0.77, 6.82]	3.55 [0.90, 7.05]	3.13 [0.55, 6.58]	-0.42 [-1.80, 0.82]
vertical fold	Qianfan-OCR	0.10 [-1.03, 1.14]	3.44 [0.91, 6.72]	3.78 [1.45, 6.89]	4.52 [2.11, 7.48]	0.75 [-0.94, 2.38]
curl	GLM-OCR	-0.01 [-0.62, 0.51]	0.44 [-0.16, 1.00]	0.56 [0.11, 1.06]	-0.11 [-0.87, 0.54]	-0.68 [-1.41, -0.11]
curl	MinerU2.5-Pro	0.05 [-0.18, 0.29]	0.22 [-0.77, 1.48]	0.35 [-0.56, 1.57]	-1.14 [-3.80, 0.34]	-1.49 [-4.22, 0.18]
curl	OvisOCR2	-0.10 [-1.37, 1.07]	-1.59 [-5.24, 1.07]	-1.24 [-4.84, 1.22]	-1.13 [-4.87, 1.42]	0.11 [-0.64, 0.91]
curl	PaddleOCR-VL-1.5	-0.19 [-0.56, 0.16]	0.19 [-0.87, 1.21]	-0.03 [-1.00, 0.85]	0.83 [-0.31, 2.09]	0.85 [0.05, 1.78]
curl	PaddleOCR-VL-1.6	-0.31 [-1.25, 0.20]	-0.39 [-1.39, 0.28]	-0.27 [-1.21, 0.35]	-0.34 [-1.37, 0.33]	-0.07 [-0.55, 0.35]
curl	HunyuanOCR	3.34 [0.75, 6.38]	5.94 [1.66, 10.54]	4.08 [-0.24, 8.76]	2.26 [-2.26, 7.15]	-1.82 [-5.76, 1.90]
curl	FireRed-OCR	-0.14 [-1.19, 0.95]	-0.15 [-1.90, 1.76]	0.60 [-1.13, 2.47]	0.52 [-0.74, 2.10]	-0.08 [-1.40, 1.62]
curl	Qianfan-OCR	-0.34 [-1.53, 0.79]	-0.88 [-2.38, 0.79]	-0.40 [-1.63, 1.10]	0.06 [-1.60, 1.92]	0.46 [-0.95, 1.87]
motion blur	GLM-OCR	16.00 [6.35, 26.89]	11.19 [0.41, 23.03]	16.82 [7.02, 27.72]	16.94 [6.89, 28.07]	0.12 [-0.73, 0.97]
motion blur	MinerU2.5-Pro	-0.15 [-0.59, 0.12]	-0.06 [-1.08, 1.12]	-0.28 [-1.55, 1.05]	-1.03 [-3.62, 0.38]	-0.74 [-3.57, 1.38]
motion blur	OvisOCR2	-0.08 [-0.28, 0.01]	-0.97 [-3.36, 0.52]	-1.06 [-3.48, 0.36]	-0.32 [-2.84, 1.43]	0.75 [-0.10, 1.88]
motion blur	PaddleOCR-VL-1.5	-0.09 [-0.77, 0.61]	0.44 [-0.31, 1.16]	0.55 [-0.34, 1.50]	1.11 [0.59, 1.69]	0.56 [-0.30, 1.37]
motion blur	PaddleOCR-VL-1.6	-0.11 [-0.22, -0.02]	-0.33 [-0.86, 0.12]	-0.41 [-0.97, 0.02]	0.17 [-0.09, 0.50]	0.58 [-0.01, 1.28]
motion blur	HunyuanOCR	2.42 [0.70, 4.24]	1.05 [-2.25, 4.28]	1.16 [-2.27, 4.40]	-0.80 [-2.91, 1.45]	-1.96 [-5.88, 1.93]
motion blur	FireRed-OCR	-0.10 [-0.46, 0.20]	-0.86 [-1.98, 0.13]	-0.28 [-1.21, 0.59]	-0.76 [-1.77, 0.12]	-0.48 [-2.03, 0.99]
motion blur	Qianfan-OCR	-0.39 [-1.42, 0.64]	-0.50 [-1.96, 0.95]	-0.75 [-2.35, 0.82]	0.43 [-0.80, 1.87]	1.18 [-0.70, 3.33]
cast shadow	GLM-OCR	0.19 [-0.45, 0.83]	0.65 [0.10, 1.29]	0.60 [0.03, 1.22]	0.36 [-0.11, 0.84]	-0.24 [-0.96, 0.33]
cast shadow	MinerU2.5-Pro	0.10 [-0.58, 1.01]	0.60 [-0.59, 2.08]	0.25 [-1.10, 1.81]	-0.77 [-2.49, 0.14]	-1.02 [-3.80, 0.97]
cast shadow	OvisOCR2	0.08 [0.00, 0.22]	0.24 [-1.05, 1.34]	-0.08 [-1.36, 0.93]	0.22 [-0.24, 0.76]	0.30 [-0.44, 1.25]
cast shadow	PaddleOCR-VL-1.5	0.16 [-0.13, 0.51]	0.24 [-0.31, 0.84]	0.03 [-0.46, 0.59]	1.09 [0.31, 1.99]	1.06 [0.26, 1.88]
cast shadow	PaddleOCR-VL-1.6	-0.07 [-0.23, 0.08]	-0.09 [-0.76, 0.55]	0.01 [-0.52, 0.66]	-0.04 [-0.18, 0.11]	-0.05 [-0.67, 0.49]
cast shadow	HunyuanOCR	-0.33 [-2.88, 2.23]	1.91 [-1.94, 5.77]	0.65 [-3.33, 4.74]	-0.79 [-2.90, 1.13]	-1.44 [-5.55, 2.75]
cast shadow	FireRed-OCR	0.43 [-0.16, 1.17]	0.31 [-0.86, 1.74]	0.41 [-0.90, 1.86]	-0.02 [-0.88, 0.87]	-0.43 [-1.85, 0.99]
cast shadow	Qianfan-OCR	-0.32 [-1.19, 0.25]	-0.03 [-1.11, 1.10]	0.06 [-0.94, 1.11]	-0.17 [-0.96, 0.58]	-0.23 [-1.52, 1.01]
table texture	GLM-OCR	-0.32 [-1.22, 0.43]	0.60 [-0.01, 1.37]	0.25 [-0.68, 1.14]	-0.20 [-1.01, 0.46]	-0.45 [-1.16, 0.09]
table texture	MinerU2.5-Pro	0.31 [-0.35, 1.26]	0.62 [-0.43, 1.95]	0.49 [-0.69, 1.89]	-0.69 [-2.41, 0.25]	-1.18 [-3.98, 0.68]
table texture	OvisOCR2	-0.06 [-1.40, 0.93]	0.22 [-2.00, 2.35]	0.81 [-0.98, 2.89]	0.59 [-1.16, 2.57]	-0.23 [-0.69, 0.08]
table texture	PaddleOCR-VL-1.5	-1.05 [-2.40, 0.03]	-0.79 [-1.92, 0.24]	-1.00 [-2.48, 0.32]	-0.47 [-1.98, 0.82]	0.53 [-0.24, 1.35]
table texture	PaddleOCR-VL-1.6	-0.62 [-1.78, 0.23]	-0.54 [-1.59, 0.32]	-0.64 [-1.86, 0.31]	-0.62 [-1.79, 0.24]	0.02 [-0.39, 0.39]
table texture	HunyuanOCR	-2.97 [-6.43, 0.14]	-0.80 [-4.18, 2.35]	-3.74 [-8.03, 0.17]	-3.81 [-7.69, -0.43]	-0.07 [-4.77, 4.67]
table texture	FireRed-OCR	0.37 [-0.47, 1.31]	0.40 [-1.37, 2.10]	1.09 [-0.02, 2.56]	0.28 [-0.67, 1.52]	-0.81 [-2.01, 0.13]
table texture	Qianfan-OCR	-0.06 [-0.56, 0.45]	-0.84 [-2.48, 0.71]	-0.95 [-2.47, 0.51]	-0.32 [-1.06, 0.29]	0.64 [-0.76, 2.09]

parser m and real condition c, define $D^{\text{real}}_{m,c} = A_m(\text{Scanning}) - A_m(c)$. For a simulated factor f, define $D^{\text{sim}}_{m,f} = A_m(N_f) - A_m(P_f)$ on the same matched support, first averaging sampled variants within each task on each page. For a factor family, we average member-factor aggregate drops with

equal weights. Kendall τ measures agreement between the eight-parser rankings by aggregate drop; larger drops mean less robustness.

The protocol maps real Skew to camera pitch, roll, and yaw; Warping to 11 paper-geometry factors; and Illumination to four illumination factors. Screen-Photography is an indirect-coverage control, compared with spotlight, ISO noise, and chromatic aberration; screen sampling artifacts such as moiré fall outside this mapping. The roll-only Skew result ($\tau = 0.86$, 95% interval $[0.57, 0.93]$) provides a factor-level sensitivity check. For each mapping, we resample source pages jointly on both sides 2,000 times (seed 20260923), recompute eight-system aggregate drops and rankings, and report percentile 95% intervals. Table 3 compares rankings by acquisition-induced score drops on the full 1,355-page matched support for the eight parsers and specified condition mappings.

We can observe that our simulated acquisition factors can capture some of the real acquisition effects. Beyond evaluation, the same physical simulation and rendering pipeline could augment OCR and document-parsing training data with varied in-the-wild acquisition conditions, potentially improving robustness to real-world acquisition changes.