OCR-Robust: How Robust is OCR-Reasoning? A Benchmark for Vision-Language Models under Visual Perturbations

Anonymous ACL submission

Abstract

Vision-language models (VLMs) achieved strong performance on OCR-based benchmarks and increasingly focused on text-rich understanding, but their robustness under realistic visual degradation remains insufficiently understood. This gap is especially critical for OCR reasoning, where visual corruption can induce OCR errors and structural distortions that propagate into downstream reasoning failures. We introduce OCR-Robust, a benchmark for evaluating OCR reasoning robustness under visual perturbations. It contains 812 samples across two complementary subsets: OCR1.0, covering documents, scene text, receipts, handwriting, and mathematical content, and OCR2.0, focusing on charts, geometry diagrams, and tables. To enable efficient yet informative evaluation, we conduct a pilot study over 18 candidate perturbations and select 5 representative types at 3 severity levels each based on their impact and cross-model discriminability. We evaluate robustness using clean accuracy, Relative Corruption Retention (RCR), Worst-Case Retention (WCR), and a composite Corruption Robustness Index (CRI), and benchmark 18 models spanning proprietary systems, open-source VLMs, and OCR+LLM pipelines. Our results show that higher clean accuracy does not necessarily imply stronger robustness, thinking-mode models suffer pronounced worst-case degradation on structure-sensitive OCR tasks, and charts and tables are substantially more fragile than document-like inputs under perturbation.

1 Introduction

Vision-Language Models (VLMs) have achieved remarkable performance on OCR and visual reasoning benchmarks. Models such as GPT-5.2 (Singh et al., 2025), Qwen3-VL (Bai et al., 2025), and Gemini-3-Flash (Google DeepMind, 2025a) demonstrate strong capabilities in document understanding, chart interpretation, and scene text recognition. Meanwhile, specialized OCR systems such as HunyuanOCR (Team et al., 2025a), PaddleOCR (Cui et al., 2025), and DeepSeekOCR (Wei et al., 2025) have advanced text extraction to near-human accuracy on clean images, approaching or surpassing human-level performance on benchmarks like OCRBench (Liu et al., 2024) and OmniDocBench (Ouyang et al., 2025). Yet on OCR-reasoning tasks that jointly demand text-rich perception and multi-step reasoning, current models still exhibit notable performance bottlenecks (Ye et al., 2025; Huang et al., 2025b). Moreover, even these evaluations are conducted on clean, well-curated images. Real-world deployment rarely presents such ideal conditions: documents are photographed under uneven lighting, receipts are crumpled and faded, and street signs are captured through rain or motion blur.

The problem extends beyond perceptual difficulty alone. OCR reasoning robustness is fundamentally different from generic visual robustness: visual corruption does not merely blur semantic content, but can induce discrete symbolic errors that are subsequently amplified by downstream reasoning. In text-rich images, even mild visual perturbations can introduce such errors: a blurred "8" becomes "3", a color-shifted decimal point disappears, or a noisy table border merges adjacent cells. Unlike general visual recognition, where a slightly degraded object can still be identified by context, OCR errors propagate through downstream reasoning in a cascading fashion. A single misread digit in an invoice invalidates an arithmetic chain; a garbled axis label in a chart reverses the conclusion of a trend analysis. This symbolic brittleness, where continuous visual degradation causes discrete reasoning failures, makes OCR reasoning a uniquely sensitive probe of robustness.

Existing benchmarks do not isolate this phenomenon. Standard OCR evaluations (Fu et al., 2024; Ouyang et al., 2025; Huang et al., 2025b)

focus on clean images. General robustness benchmarks such as ImageNet-C (Hendrycks and Dietterich, 2019), VLM-RobustBench (Saxena et al., 2026) assess object recognition under corruption but do not involve text extraction or reasoning. Even recent OCR-centric benchmarks such as CC-OCR (Yang et al., 2025) do not isolate how visual degradation causes OCR errors that propagate through downstream reasoning. To systematically study this problem, we introduce **OCR-Robust**, a benchmark designed for evaluating OCR reasoning robustness under visual perturbations. Code and supplementary materials are available at `https://anonymous.4open.science/r/OCR-Robust-EBA4`. Our Our benchmark makes three contributions:

- **A diverse, two-part dataset.** We construct 812 samples spanning two complementary subsets: OCR1.0 (482 samples) covers documents, scene text, receipts, handwriting, screenshots, and mathematical content; OCR2.0 (330 samples) targets charts, geometry diagrams, and tables. Together they cover the full spectrum from natural text perception to structured visual reasoning.

- **A principled perturbation selection pipeline.** Rather than applying perturbations ad hoc, we conduct a pilot study evaluating 18 candidate perturbation types across 6 models, using discriminability metrics to select the 5 perturbations that most effectively differentiate model robustness: glass blur, motion blur, elastic deformation, color shift, and snow. Each is applied at 3 severity levels, yielding 15 evaluation conditions.

- **A robustness metric suite.** We propose four complementary metrics: Clean Accuracy, Relative Corruption Retention (RCR), Worst-Case Retention (WCR), and a composite Corruption Robustness Index (CRI) that balances baseline capability, average retention, and worst-case risk through geometric mean aggregation.

Through extensive evaluation of 18 models spanning proprietary systems (GPT-5.2 (Singh et al., 2025), Gemini-3-Flash (Google DeepMind, 2025a)), open-source VLMs (Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025b)), and OCR+LLM pipelines (PaddleOCR-VL-1.5 (Cui et al., 2026) + GPT-5.1), we observe five notable findings: (1) closed-source models tend to lead in overall robustness, with Gemini-3-Flash achieving CRI above 0.90 on OCR1.0; (2) higher capability does not necessarily translate to stronger robustness—thinking-mode models achieve the highest clean accuracy but can suffer sharp worst-case drops on structured content (WCR as low as 0.579); (3) structured visual content (charts, tables) appears substantially more fragile than documents and scene text, with average worst-case retention declining from 0.826 on OCR1.0 to 0.676 on OCR2.0; (4) OCR+LLM pipeline performance appears limited by extraction errors rather than reasoning quality; (5) prompt-induced chain-of-thought (CoT) improves clean accuracy but does not reliably prevent worst-case failures, suggesting that symbolic brittleness may be most pronounced where reasoning demands are highest.

2 Related Work

2.1 OCR and Text-Rich Multimodal Benchmarks

Modern VLMs are expected not only to recognize text but also to localize, parse layouts, and reason over textual evidence. This has driven the development of comprehensive OCR benchmarks. OCRBench (Liu et al., 2024) and its successor OCRBench v2 (Fu et al., 2024) cover tasks from basic recognition to complex visual text reasoning. CC-OCR (Yang et al., 2025) targets multilingual reading and document parsing, while OmniDocBench (Ouyang et al., 2025) focuses on detailed PDF understanding. For long documents, MMLongBench (Wang et al., 2025c) and M-LongDoc (Chia et al., 2025) require cross-page evidence integration. More recently, benchmarks have shifted toward OCR reasoning: LogicOCR (Ye et al., 2025) tests logical inference over text-rich images, while OCR-Reasoning (Huang et al., 2025b) categorizes evaluations by fine-grained reasoning abilities.

2.2 Robustness Evaluation for VLMs

Recent work has increasingly examined model stability under naturalistic corruptions. Ishmam et al. (2025) stress-tests VQA models with 213K augmented images and proposes aggregatable robustness metrics. MLLM-IC (Qiu et al., 2025) evaluates models across 40 corruption types and 34 capabilities for fine-grained diagnosis. Besides comprehensive variations types (Saxena et al., 2026;

Fan et al., 2025), other benchmarks address multi-modal shifts (Huang et al., 2025a), self-consistency under corruption (Zhang et al., 2024), resolution robustness (Li et al., 2025), and coordinated cross-modal perturbations (Babu et al., 2025).

3 Base Benchmark Construction

We construct two complementary datasets combining multiple sources to ensure comprehensive coverage of OCR reasoning scenarios. Our approach balances efficiency by leveraging existing high quality datasets, coverage through targeted annotation, and quality via human review of VLM generated questions. These datasets serve as the base set for our robustness evaluation.

3.1 Data Sources and Composition

Our benchmark comprises 812 samples across two complementary datasets, as summarized in Table 1. OCR1.0 (482 samples) spans nine data types covering documents, scene text, receipts, handwriting, and mathematical content, focusing on multi-hop reasoning and cross-reference integration. OCR2.0 (330 samples) draws from ChartVQA (Masry et al., 2022), GNS-260K (Ning et al., 2025), and TableVQA-Bench (Kim et al., 2024), targeting chart, geometry, and table comprehension that requires analysis of structured visual content, such as trend analysis, value aggregation, and spatial reasoning. Together, the two subsets cover the full spectrum from document-level text extraction to structured visual reasoning.

Dataset	Data Type	Source	#
OCR 1.0	Eng. Document	DocVQA	50
	Scene Text	TextVQA	52
	Infographic	InfographicVQA	40
	Chn. Advertisement	MTWI	63
	Receipt	SROIE	75
	Screenshot	Screen2Words	30
	Handwritten	OCR-Reasoning	30
	Mathematical	GSM8K/TheoremQA	100
	Chn. Document	M6Doc	42
OCR 2.0	Chart	ChartVQA	150
	Geometry	GNS-260K	80
	Table	TableVQA-Bench	100
Total			812

Table 1: Distribution of data types in the benchmark.

3.2 Construction Methods

We employ three complementary approaches tailored to different data characteristics:

Direct Selection For OCR2.0 and portions of OCR1.0 (handwriting from OCR-Reasoning (Huang et al., 2025b) and InfographicVQA (Mathew et al., 2021) subsets), we select samples from existing high-quality datasets. OCR2.0 employs visual diversity sampling to ensure broad coverage of chart types. We extract 2048-dimensional ResNet50 (He et al., 2015) features from candidate images, apply agglomerative clustering (Müllner, 2011) with Ward linkage to form 150 clusters, and select the sample closest to each cluster centroid. This automated approach efficiently identifies diverse samples while preserving original questions and answers.

VLM-Assisted Annotation For OCR1.0, the DocVQA (Mathew et al., 2020), TextVQA (Singh et al., 2019), MTWI (He et al., 2018), SROIE (Huang et al., 2019), Screen2Words (Wang et al., 2021), and M6Doc (Cheng et al., 2023) subsets undergo VLM-assisted annotation to generate reasoning questions. We use three state-of-the-art VLMs: GPT-5.2 (Singh et al., 2025), Gemini-3-Pro (Google DeepMind, 2025b), and Claude-Sonnet-4.5 (Anthropic, 2025). For each image, one randomly selected model generates a reasoning question and provides the corresponding answer. Using multiple models also helps mitigate potential data contamination issues. All generated question-answer pairs are subsequently reviewed and verified by human annotators to ensure quality and correctness.

Synthetic Generation The math subset in OCR1.0 is synthetically constructed from GSM8K (Cobbe et al., 2021) and TheoremQA (Chen et al., 2023). Following the LogicOCR (Ye et al., 2025) pipeline, we first convert multiple-choice questions into free-form answer format to better evaluate reasoning ability. Then we use GPT-4o to generate a visual context description for each question, and leverage GPT-4o-image to produce the corresponding image with the context, question, and answer rendered into it. This pipeline naturally provides ground-truth answers, and we further use PaddleOCR (Cui et al., 2025) to compute character recall for filtering.

3.3 Quality Control

We enforce multiple quality control criteria throughout dataset construction. First, we manually verify both answer verifiability and OCR necessity,

ensuring that each sample has an objectively supported answer grounded in the image content and cannot be solved without text extraction. Second, we perform near-duplicate removal based on perceptual image hashing. Finally, we perform a final manual review of annotated samples to ensure correctness. Only samples that pass all quality criteria are retained in the final benchmark.

4 Experiments

4.1 Experimental Setup

Evaluated Models We evaluate three categories of systems on OCR-Robust: closed-source VLMs, open-source VLMs, and OCR+LLM pipelines.

(a) Open-source MLLMs. We assess a broad range of open-source models spanning diverse scales and architectures: Qwen2.5-VL-7B (Team, 2025), Qwen3-VL-4B/8B/32B-Instruct (Bai et al., 2025), Qwen3-VL-4B/8B-Thinking (Bai et al., 2025), InternVL3.5-4B/8B/14B/38B (Wang et al., 2025b), Llama-4-Scout-17B-16E-Instruct (Meta, 2025), Kimi-VL-A3B-Thinking (Team et al., 2025b), VL-Rethinker-7B (Wang et al., 2025a), and MM-Eureka-Qwen-7B (Meng et al., 2025). Among these, the Thinking variants employ extended chain-of-thought reasoning during inference. **(b) Closed-source MLLMs.** We also evaluate leading closed-source models, including GPT-5.2 (Singh et al., 2025), GPT-5.1 (Singh et al., 2025), and Gemini-3-Flash (Google DeepMind, 2025a) (with and without thinking mode). **(c) OCR + LLM Pipelines.** Finally, to examine whether decoupling perception from reasoning offers robustness advantages, we evaluate a two-stage pipeline where PaddleOCR-VL-1.5 (Cui et al., 2026) first extracts text from images, which is then fed to a LLM for text-only reasoning. We test GPT-5.1 (Singh et al., 2025) and Gemini-3-Flash (thinking mode) (Google DeepMind, 2025a) as downstream reasoning models.

Implementation Details We adopt a zero-shot evaluation protocol without fine-tuning or few-shot prompting. For VLMs, we present both the image and the textual question, appended with a format hint from the dataset annotation (e.g., "The composition of the final answer should be: $ + Integer") to ensure a consistent output format. The prompt ends with "Directly output the answer only, without any explanation." For OCR + LLM pipelines, we replace the image input with text extracted by PaddleOCR-VL-1.5 while keeping the same query structure. Scoring uses normalized exact match, with GPT-4o as a semantic fallback judge, as detailed in Appendix A.3.

Metrics We evaluate robustness from four complementary perspectives. Let $A^{\text{clean}}_{m,d}$ denote the accuracy of model m on dataset d under clean images, and $A_{m,d,c,s}$ denote the accuracy under corruption family c at severity level s. Let $\Omega_d = \{(c, s)\}$ be the set of all observed perturbation conditions on dataset d.

Clean Accuracy is the baseline accuracy on unperturbed images, reported separately to distinguish inherent capability from robustness.

Relative Corruption Retention (RCR) measures the average ratio of corrupted accuracy to clean accuracy:

$$\text{RCR}(m, d) = \frac{1}{|\Omega_d|} \sum_{(c,s)\in\Omega_d} \frac{A_{m,d,c,s}}{A^{\text{clean}}_{m,d}} \quad (1)$$

A value of 1 indicates perfect retention. Unlike mean corruption accuracy (mCA), RCR normalizes by baseline performance, enabling fair comparison between models of different capability levels.

Worst-Case Retention (WCR) captures the most severe failure mode by normalizing the minimum corrupted accuracy by the clean baseline:

$$\text{WCR}(m, d) = \frac{\min_{(c,s)\in\Omega_d} A_{m,d,c,s}}{A^{\text{clean}}_{m,d}} \quad (2)$$

A model with high RCR but low WCR has a hidden vulnerability in specific corruption types.

Corruption Robustness Index (CRI) balances clean capability, average retention, and worst-case risk via the geometric mean of three factors:

$$\text{CRI}(m, d) = \sqrt[3]{\frac{A^{\text{clean}}_{m,d}}{100} \cdot \text{RCR}(m, d) \cdot \text{WCR}(m, d)} \quad (3)$$

The geometric mean (cube root of the product) penalizes imbalance: a model that excels in one dimension but fails in another receives a lower CRI than one that performs consistently across all three, preventing high baseline accuracy from masking poor robustness. $A^{\text{clean}}_{m,d}$ is divided by 100 for normalization.

4.2 Perturbation Selection

Not all visual perturbations are equally suitable for OCR reasoning robustness evaluation. A useful perturbation should jointly satisfy several criteria:

it should meaningfully affect performance, distinguish models with different robustness levels, induce monotonic degradation as severity increases, complement rather than duplicate other perturbations, and reflect distinct real-world degradation mechanisms. We therefore conduct a pilot study over 18 candidate perturbations across multiple VLMs to identify an informative subset.

Experimental Setup We evaluated 6 VLMs covering closed-source (GPT-5.2), open-source (InternVL3.5-8B, Qwen2.5-VL-7B, Qwen3-VL-4B/8B-Instruct), and thinking-mode (Qwen3-VL-8B-Thinking) paradigms across **18 candidate perturbation types** in 5 categories: Noise (Gaussian, shot, impulse), Blur (defocus, glass, motion, zoom), Weather (fog, frost, snow), Digital (brightness, contrast, elastic, pixelate), and Color & Geometric (color shift, glare, moiré, slight rotation). Each perturbation was applied at 4 severity levels controlled by LPIPS (Zhang et al., 2018) distance (0.05, 0.10, 0.15, 0.20), ensuring perceptually uniform degradation. We used a stratified sample of 100 questions from the OCR-Reasoning dataset (Huang et al., 2025b), covering diverse visual contexts.

Selection Metrics We evaluate each candidate perturbation along three complementary axes. Let $r_{m,c,s} = A_{m,c,s}/A_m^{\text{clean}}$ denote the retention ratio of model m under perturbation c at severity s, with M models and S_c severity levels per perturbation. **Mean Relative Drop (MRD)** measures the average performance degradation a perturbation induces across all models and severity levels:

$$\text{MRD}(c) = \frac{1}{M \cdot |S_c|} \sum_{m,s} (1 - r_{m,c,s}) \tag{4}$$

Higher MRD indicates a stronger perturbation. However, high impact alone is insufficient: a perturbation is not informative if it degrades all models in nearly the same way.

Separability (SEP) quantifies how effectively a perturbation differentiates models with distinct robustness profiles, measured as the average pairwise retention divergence:

$$\text{SEP}(c) = \frac{1}{|S_c|} \sum_{s \in S_c} \frac{2}{M(M-1)} \sum_{i<j} |r_{i,c,s} - r_{j,c,s}| \tag{5}$$

High SEP ensures the perturbation reveals meaningful differences between models rather than affecting all models equally.

Monotonicity (MON) measures whether performance degrades consistently as severity increases, rather than fluctuating erratically:

$$\text{MON}(c) = 1 - \frac{1}{M(|S_c|-1)} \sum_m \sum_s \mathbf{1}[r_{m,c,s+1} > r_{m,c,s}] \tag{6}$$

A perturbation with high MON produces a reliable severity gradient, enabling fine-grained robustness profiling.

Results Table 2 and Figure 1 present the three selection metrics for all 18 candidates. As Figure 1 illustrates, several alternative perturbations are reasonably competitive. Rather than selecting perturbations based on a single ranking, we first filter out candidates that catastrophically destroy text information rather than degrading it gradually. For example, pixelate is excluded despite its highest MRD of 0.461, as it renders characters unrecognizable at moderate severity and makes the task effectively unsolvable. We then apply greedy selection to the remaining candidates, selecting five perturbations in order: **glass blur** (MRD = .445, SEP = .258, MON = .667) as the highest-impact non-destructive perturbation, **motion blur** (.292, .313, .611) for its strong separability, **elastic deformation** (.310, .211, .889) for its high monotonicity and low redundancy with blur types, **color shift** (.230, .112, .833) to cover color-based degradation, and **snow** (.217, .211, .778) to represent weather-induced corruption. The final subset spans four distinct corruption families while maintaining reasonable scores across all three axes. We note that alternative subsets with comparable overall quality are possible; our selection prioritizes category diversity and metric balance rather than claiming unique optimality. Bootstrap stability analysis (1,000 resamples, Appendix A.4) shows that these five perturbations are among the most frequently chosen, with selection rates ranging from 44.2% to 98.4%, though the moderate rate for color shift (44.2%) indicates that substitution with other color-category perturbations could yield a comparably effective subset.

Final Perturbation Configuration Each selected perturbation is configured with 3 severity levels: **Glass blur** (perturbation radius: 2, 2.5, 3 pixels), **Color shift** (channel offset: 3, 4, 5 pixels), **Elastic transform** (deformation strength α: 10, 15, 20), **Motion blur** (kernel size: 5, 6, 7 pixels), and **Snow** (layer intensity: 0.1, 0.2, 0.3). These parameters are calibrated to produce perceptually mean-

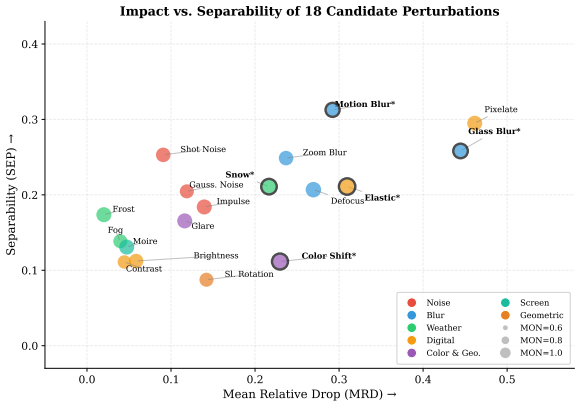

Figure 1: Impact (MRD) vs. Separability (SEP) for 18 candidate perturbations. Point size encodes Monotonicity (MON), color indicates corruption category. Black circles mark the 5 selected perturbations.

Perturbation	MRD↑	SEP↑	MON↑	Cat.
Pixelate	.461	.295	.722	Digital
Glass Blur*	.445	.258	.667	Blur
Elastic*	.310	.211	.889	Digital
Motion Blur*	.292	.313	.611	Blur
Defocus Blur	.269	.207	.778	Blur
Zoom Blur	.237	.249	.667	Blur
Color Shift*	.230	.112	.833	Color
Snow*	.217	.211	.778	Weather
Slight Rotation	.142	.087	.611	Geometric
Impulse Noise	.140	.184	.722	Noise
Gaussian Noise	.119	.205	.611	Noise
Glare	.116	.165	.722	Color
Shot Noise	.091	.253	.667	Noise
Brightness	.058	.112	.667	Digital
Moiré	.047	.131	.722	Screen
Contrast	.044	.111	.556	Digital
Fog	.040	.138	.611	Weather
Frost	.020	.174	.722	Weather

Table 2: Perturbation selection metrics across 6 pilot models. * denotes the final 5 perturbations.

ingful degradation while maintaining text legibility at lower severity levels, yielding 5 dimensions $\times$ 3 levels = 15 perturbation conditions.

4.3 Results and Analysis

Table 3 presents the robustness evaluation results on both OCR1.0 and OCR2.0. Models are grouped by family and ranked by average CRI.

Finding 1: Closed-source models lead in overall robustness. As shown in Figure 12, the top-performing closed-source models occupy the upper-right quadrant, achieving CRI above 0.90 on OCR1.0 with RCR $\geq$ 0.986 and WCR $\geq$ 0.920, and lead on OCR2.0 with CRI above 0.84. These models maintain near-perfect average retention while exhibiting the strongest worst-case resilience, indicating fundamentally more robust visual text understanding rather than merely higher baseline accuracy. Among open-source models, Qwen3-VL-8B-Thinking achieves the closest CRI (0.875 on OCR1.0), but its worst-case retention on OCR2.0 (0.624) reveals a significant gap compared to closed-source systems.

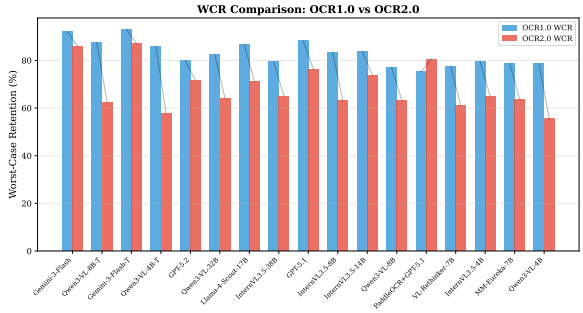

Figure 2: WCR comparison between OCR1.0 and OCR2.0. All models show lower worst-case retention on OCR2.0, confirming that structured visual content is more fragile under perturbation.

Finding 2: Higher capability does not guarantee higher robustness. This pattern manifests in two dimensions (see also the collapse curves in Figures 3 and 4). First, *native thinking models improve clean accuracy but not robustness*: Qwen3-VL-8B-Thinking, which employs internalized chain-of-thought learned during training, achieves the highest clean accuracy on OCR2.0 (78.18%) yet its WCR drops to 0.624, while Gemini-3-Flash at 71.52% clean maintains WCR of 0.860. The same Thinking models show much milder drops on OCR1.0 (WCR $\geq$ 0.859), suggesting that internalized reasoning is more vulnerable on structured visual content. Second, *larger scale does not monotonically improve robustness*: within the InternVL3.5 family, the 14B model (WCR = 0.838 on OCR1.0, 0.738 on OCR2.0) consistently outperforms the 38B variant (WCR = 0.794 and 0.650) in worst-case retention despite lower clean accuracy. This suggests that larger models may overfit to clean image distributions, becoming more brittle under perturbation.

Finding 3: Structured visual content is substantially more fragile. Across all models, WCR on OCR2.0 is consistently lower than on OCR1.0 (Figure 2). The average WCR drops from 0.826 (OCR1.0) to 0.676 (OCR2.0), a relative decline of 18%. Color shift emerges as one of the main drivers of this gap (Figure 5), with several models losing over 40% of their clean accuracy at the high-

Model	OCR1.0				OCR2.0			
	Clean	RCR↑	WCR↑	CRI↑	Clean	RCR↑	WCR↑	CRI↑
Closed-source Models								
Gemini-3-Flash-Thinking	79.05	**.998**	**.932**	**.902**	69.70	.977	**.870**	.840
Gemini-3-Flash	**80.71**	.986	.920	.901	71.52	**.980**	.860	**.845**
GPT-5.2	74.69	.911	.799	.816	**71.82**	.876	.717	.767
GPT-5.1	60.58	.986	.882	.808	49.70	.931	.762	.706
Open-source Models								
Qwen3-VL-8B-Thinking	79.67	.961	.875	.875	**78.18**	.892	.624	.758
Qwen3-VL-4B-Thinking	78.63	.948	.859	.862	76.97	.879	.579	.732
Qwen3-VL-32B-Instruct	72.82	.954	.826	.831	57.27	.928	.640	.698
Qwen3-VL-8B-Instruct	54.15	.903	.770	.722	49.39	.921	.632	.660
Qwen3-VL-4B-Instruct	49.79	.927	.788	.714	49.70	.841	.555	.615
InternVL3.5-38B	62.86	.933	.794	.775	61.52	.941	.650	.722
InternVL3.5-14B	55.81	.939	.838	.760	55.45	.912	.738	.720
InternVL3.5-8B	57.26	.941	.835	.766	56.67	.892	.631	.683
InternVL3.5-4B	52.49	.935	.796	.731	51.52	.905	.647	.671
Llama-4-Scout-17B	67.84	.939	.867	.821	59.70	.866	.711	.716
MM-Eureka-Qwen-7B	51.25	.918	.789	.719	47.58	.888	.637	.646
VL-Rethinker-7B	52.49	.914	.776	.719	48.18	.928	.610	.649
OCR + LLM Pipelines								
PaddleOCR-VL-1.5 + GPT-5.1	52.90	.902	.753	.711	26.36	.982	.805	.593
PaddleOCR-VL-1.5 + Gemini-3-Flash-T	65.15	.938	.796	.787	48.79	.910	.665	.666

Table 3: Robustness evaluation on OCR1.0 (482 samples) and OCR2.0 (330 samples). Models grouped by family, ranked by average CRI. Best values per dataset in **bold**.

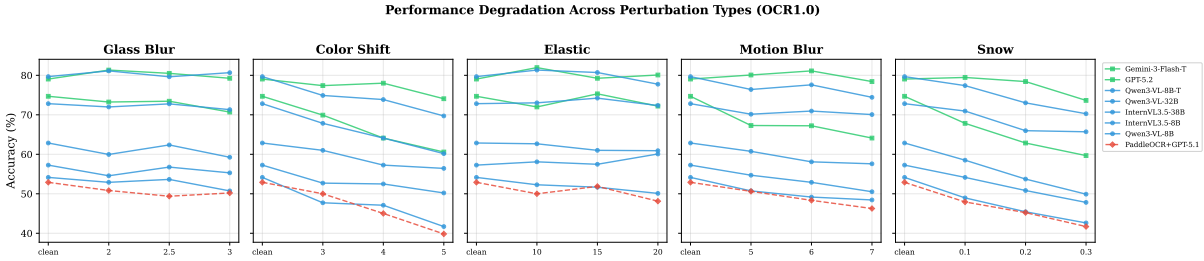

Figure 3: Performance degradation across five perturbation types on OCR1.0. Each subplot shows accuracy from clean (leftmost) to highest severity (rightmost). Closed-source models (green squares) maintain flatter curves, while pipelines (red diamonds) exhibit steeper drops.

est severity on OCR2.0. This suggests that charts, tables, and geometry diagrams are more sensitive to visual perturbations than natural scene text or documents, likely because structured visual content often depends on fine-grained visual cues for element differentiation and relational interpretation.

Finding 4: Pipeline robustness is bottlenecked by OCR extraction. The OCR + LLM pipelines exhibit a characteristic trade-off. PaddleOCR-VL-1.5 + GPT-5.1 achieves high RCR (0.902 on OCR1.0, 0.982 on OCR2.0) but the lowest CRI due to severely limited clean accuracy (52.90% and 26.36%). Replacing the reasoning backend with Gemini-3-Flash-Thinking improves clean accuracy substantially (65.15% and 48.79%) and CRI (0.787 and 0.666), but WCR remains comparable (0.796 and 0.665). This suggests that robustness in the two-stage pipeline is primarily constrained by the OCR extraction stage rather than the downstream reasoning model: stronger reasoning improves the clean-performance ceiling, but offers limited protection once perturbation-induced OCR errors have already corrupted the symbolic input.

Finding 5: Prompt-induced CoT improves accuracy but not worst-case robustness. Table 4 compares standard (direct-answer) and prompt-induced CoT on three open-source *non-thinking* models (i.e., models without built-in chain-of-

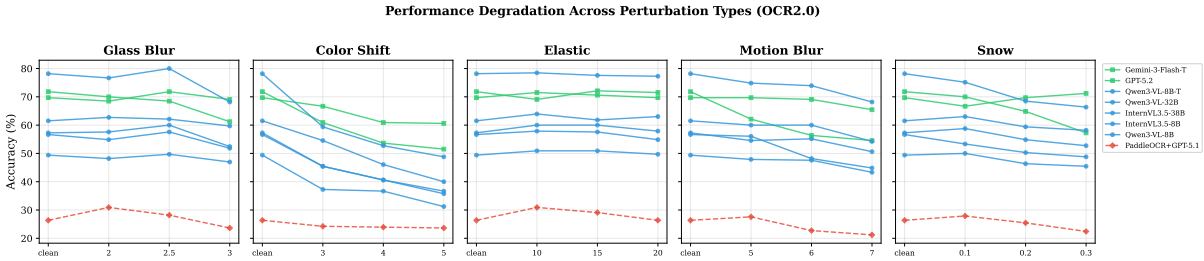

Figure 4: Performance degradation across five perturbation types on OCR2.0. Compared to OCR1.0, all models show sharper collapses, particularly under color shift and snow perturbations.

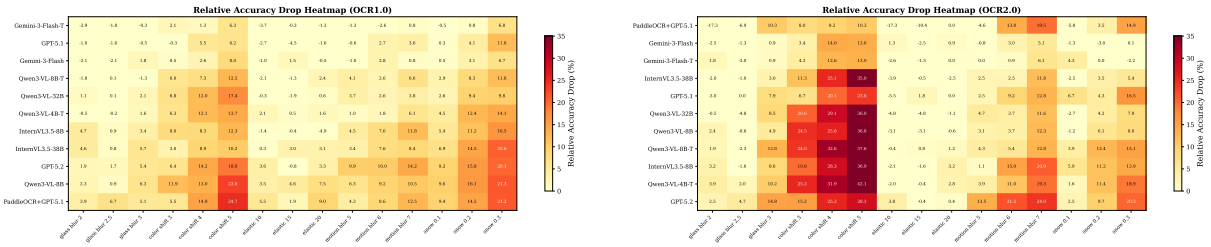

Figure 5: Relative accuracy drop (%) heatmap across all perturbation conditions. Darker cells indicate larger drops. Snow and color shift at high severity cause the most severe degradation, especially on OCR2.0.

Model	Prompt	Clean	RCR	WCR	CRI
		OCR1.0			
Qwen3-VL-8B	Std	54.15	.903	.770	.722
	CoT	79.25	.965	.874	.875
Qwen3-VL-4B	Std	49.79	.927	.788	.714
	CoT	79.67	.942	.814	.849
InternVL3.5-8B	Std	57.26	.941	.835	.766
	CoT	74.69	.929	.861	.842
		OCR2.0			
Qwen3-VL-8B	Std	49.39	.921	.632	.660
	CoT	76.06	.893	.594	.739
Qwen3-VL-4B	Std	49.70	.841	.555	.615
	CoT	70.91	.910	.624	.738
InternVL3.5-8B	Std	56.67	.892	.631	.683
	CoT	61.82	.893	.652	.711

Table 4: Effect of CoT prompting on robustness.

thought training). This isolates the effect of prompting strategy from architectural differences in native thinking models discussed in Finding 2. On OCR1.0, CoT consistently improves all metrics: clean accuracy rises by 17–30 points, and CRI improves by 0.08–0.15. However, on OCR2.0 the picture diverges. While CoT still boosts clean accuracy substantially (+5 to +27), worst-case retention shows minimal improvement or even decline: Qwen3-VL-8B with CoT drops WCR from 0.632 to 0.594. Combined with Finding 2, both forms of extended reasoning, whether internalized during training (Thinking models) or elicited through prompting (CoT), amplify clean performance but fail to shield against the worst perturbation scenarios on structured visual content, where corrupted symbolic inputs lead reasoning chains down incorrect paths.

5 Conclusion

We present OCR-Robust, a benchmark for evaluating VLMs robustness on OCR reasoning tasks under visual perturbations. Evaluation of 18 models across 812 samples and 5 perturbation types reveals five findings: (1) closed-source models lead in overall robustness, with Gemini-3-Flash achieving CRI above 0.90; (2) higher clean accuracy does not guarantee robustness—thinking-mode models and larger variants often exhibit worse worst-case retention; (3) structured visual content (charts, tables) is substantially more fragile than documents and scene text; (4) OCR+LLM pipelines are bottlenecked by extraction errors rather than reasoning quality; (5) prompt-induced chain-of-thought improves clean accuracy but fails to prevent worst-case failures on corrupted inputs. These findings highlight a critical gap between benchmark performance and real-world reliability. We hope OCR-Robust encourages the development of VLMs that are both accurate and resilient to visual degradations encountered in practical deployment.

Limitations

The dataset (812 samples) is moderate in size and language coverage. The perturbation selection pilot study uses a narrow basis (100 questions, 6 models, single source dataset), and alternative perturbation subsets with comparable diagnostic value are plausible. We evaluate 5 perturbation types individually, whereas real world degradations often co-occur; compound perturbations remain future work. Our zero-shot evaluation protocol may underestimate models' robustness under optimized prompting. The pipeline evaluation is limited to PaddleOCR-VL-1.5 as the sole OCR engine. Due to budget constraints, closed-source coverage is limited to GPT-5.x and Gemini-3-Flash. Finally, although all annotations undergo human review, this introduces a potential data contamination risk.

References

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Ashwin Ramesh Babu, Sajad Mousavi, Vineet Gundecha, Sahand Ghorbanpour, Avisek Naug, Antonio Guillen, Ricardo Luna, and Soumyendu Sarkar. 2025. Coordinated robustness evaluation framework for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 712–720.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 2023. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15138–15147.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2025. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9244–9261.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

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. 2026. Paddleocr-vl-1.5: Towards a multi-task 0.9b vlm for robust in-the-wild document parsing. *Preprint*, arXiv:2601.21957.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. 2025. Unveiling the lack of LVLM robustness to fundamental visual variations: Why and path forward. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20222–20242, Vienna, Austria. Association for Computational Linguistics.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, and 1 others. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.

Google DeepMind. 2025a. Gemini 3 flash. https://deepmind.google/models/gemini/flash/.

Google DeepMind. 2025b. Gemini 3 pro. https://deepmind.google/models/gemini/pro/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. 2018. Icpr2018 contest on robust reading for multi-type web images. pages 7–12.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*.

Chengyue Huang, Brisa Maneechotesuwan, Shivang Chopra, and Zsolt Kira. 2025a. Frames-vqa: Benchmarking fine-tuning robustness across multi-modal shifts in visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3909–3918.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2025b. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, page 1516–1520. IEEE.

Md Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, Abu Raihan Mostofa Kamal, and Md Azam Hossain. 2025. Visual robustness benchmark for visual question answering (vqa). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6623–6633. IEEE.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Chenxu Li, Zhicai Wang, Yuan Sheng, Xingyu Zhu, Yanbin Hao, and Xiang Wang. 2025. Res-bench: Benchmarking the robustness of multimodal large language models to dynamic resolution input. *arXiv preprint arXiv:2510.16926*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021. Infographicvqa. *Preprint*, arXiv:2104.12756.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Meta. 2025. Llama-4-scout-17b-16e-instruct. `https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct`. Hugging Face model card. Model release date: April 5, 2025. Accessed: 2026-03-16.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *Preprint*, arXiv:1109.2378.

Maizhen Ning, Zihao Zhou, Qiufeng Wang, Xiaowei Huang, and Kaizhu Huang. 2025. Gns: Solving plane geometry problems by neural-symbolic reasoning with multi-modal llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24957–24965.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. 2025. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9014–9023.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. 2026. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *Preprint*, arXiv:2603.06148.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

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, and 13 others. 2025a. Hunyuanocr technical report. *Preprint*, arXiv:2511.19575.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 73 others. 2025b. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Qwen Team. 2025. Qwen2.5-vl.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2words: Automatic mobile ui summarization with multimodal learning. *Preprint*, arXiv:2108.03353.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, and 1 others. 2025c. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. *arXiv preprint arXiv:2505.10610*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, and 1 others. 2025. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21744–21754.

Maoyuan Ye, Haibin He, Qihuang Zhong, Jing Zhang, Juhua Liu, and Bo Du. 2025. Logicocr: Do your large multimodal models excel at logical reasoning on text-rich images? *arXiv preprint arXiv:2505.12307*.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024. Benchmarking large multimodal models against common corruptions. *Preprint*, arXiv:2401.11943.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

A Details on Experimental Setup

A.1 Perturbation Implementation

Each of the five selected perturbations is applied at three severity levels, calibrated to produce perceptually meaningful degradation while maintaining text legibility at lower levels. All perturbations are implemented using OpenCV and PIL, with random seeds fixed per sample for reproducibility.

Selected Perturbations for Final Benchmark

Glass Blur: Simulates the effect of viewing text through textured glass. We apply a random displacement field generated by Gaussian-filtered noise, followed by Gaussian blur. Severity levels use perturbation radius: 2, 2.5, 3 pixels.

Color Shift: Introduces chromatic aberration by shifting RGB channels independently. Each channel is translated by a random offset in a random direction. Severity levels use channel offset: 3, 4, 5 pixels.

Elastic Transform: Applies smooth geometric deformation via displacement fields generated from Gaussian-filtered random noise. Deformation strength α controls the magnitude of displacement. Severity levels: $\alpha = 10, 15, 20$.

Motion Blur: Simulates camera motion during capture by convolving with a linear motion kernel at a random angle. Severity levels use kernel size: 5, 6, 7 pixels.

Snow: Overlays semi-transparent snowflake patterns generated from random noise, simulating weather-induced occlusion. Severity levels use layer intensity: 0.1, 0.2, 0.3.

All perturbations preserve the original image dimensions and format. Perturbed images are saved in PNG format to avoid compression artifacts.

Additional Perturbations Evaluated in Pilot Study The following 13 perturbations were evaluated but not selected for the final benchmark:

Noise Family: *Gaussian Noise* adds pixel-wise Gaussian noise with standard deviation $\sigma = 15, 25, 35$. *Shot Noise* simulates photon counting noise via Poisson distribution with $\lambda = 20, 40, 60$. *Impulse Noise* randomly replaces pixels with black or white (salt-and-pepper) at rates 0.01, 0.02, 0.03.

Blur Family: *Defocus Blur* applies disk-shaped blur kernels with radius 2, 3, 4 pixels. *Zoom Blur* simulates radial motion blur from camera zoom with strength 1.02, 1.04, 1.06.

Weather Family: *Fog* overlays semi-transparent white noise layers with opacity 0.3, 0.5, 0.7. *Frost*

applies crystalline patterns via Perlin noise with intensity 0.2, 0.4, 0.6.

Digital Family: *Brightness* adjusts luminance by factors 0.7, 0.5, 0.3 (darkening). *Contrast* reduces contrast by factors 0.7, 0.5, 0.3. *Pixelate* downsamples and upsamples with block sizes 4, 6, 8 pixels.

Color & Geometric Family: *Glare* overlays elliptical gradient patterns simulating light reflection with intensity 0.4, 0.6, 0.8. *Moiré* applies sinusoidal interference patterns with frequency 0.1, 0.15, 0.2. *Slight Rotation* rotates images by 2, 4, 6 degrees with border padding.

LPIPS-Controlled Severity Calibration In the pilot study, perturbation severity was controlled using LPIPS (Zhang et al., 2018) (Learned Perceptual Image Patch Similarity) distance to ensure perceptually uniform degradation across perturbation types. For each perturbation, we performed binary search over parameter ranges to find settings yielding target LPIPS distances of 0.05, 0.10, 0.15, 0.20 from the clean image. This approach ensures that severity levels are comparable across different perturbation families. For the final benchmark, we use fixed parameter values (reported above) that approximate these LPIPS targets while maintaining computational efficiency.

A.2 Prompt Templates

All models receive identical prompts with four components: (1) the input image (for MLLMs) or extracted text (for pipelines), (2) the question text, (3) a format hint specifying the expected answer structure, and (4) an instruction to output only the final answer without explanation.

Prompt Template for MLLMs:

[Image]

Question: {question_text}

{format_hint}

Directly output the answer only, without any explanation.

Figure 6: Evaluation prompt template for multimodal language models.

A.3 Evaluation Protocol

Answer Normalization Before matching, both ground-truth and predicted answers undergo normalization: (1) convert to lowercase, (2) remove punctuation except decimal points and currency symbols, (3) strip leading/trailing whitespace, (4) collapse multiple spaces to single space.

Exact Match with GPT-4o Judge Fallback We first attempt normalized exact match. If this fails and the question type is not multiple-choice, we invoke GPT-4o as a semantic judge. The judge is used for approximately 15–20% of samples across the benchmark, primarily for free-form answers where phrasing varies but meaning is preserved.

Format Hints Each question includes a format hint derived from the answer type (e.g., "The composition of the final answer should be: $ + Integer" for monetary values). These hints are appended uniformly to all models' prompts. While this may favor models trained on structured output formats, we observe no systematic bias: models with similar architectures (e.g., Qwen3-VL-4B vs. 8B) show performance gaps larger than hint-induced variance, and closed-source models maintain advantages across all answer types.

Example 1: Monetary Value

Question: What is the total amount on this receipt?

The composition of the final answer should be: $ + Integer

Directly output the answer only, without any explanation.

Figure 7: Example prompt for monetary value questions.

Example 2: Free-form Answer

Question: What is the main topic of this document?

Answer in English.

Directly output the answer only, without any explanation.

Figure 8: Example prompt for free-form questions.

A.4 Bootstrap Stability Analysis

To verify that our perturbation selection is not an artifact of the specific pilot sample, we perform a bootstrap stability analysis. We resample the 100 pilot questions with replacement 1,000 times. For each bootstrap sample, we recompute MRD, SEP,

Example 3: Numerical Calculation

Question: Based on the chart, what is the percentage increase from 2020 to 2021?

The composition of the final answer should be: Integer + %

Directly output the answer only, without any explanation.

Figure 9: Example prompt for numerical calculation questions.

GPT-4o Judge Prompt:

Question: {question}
Ground truth: {gt_answer}
Prediction: {pred_answer}

Are these two answers semantically equivalent? Answer only "yes" or "no".

Figure 10: Prompt template for GPT-4o semantic judge.

and MON for all 18 perturbations and re-run the greedy selection procedure. Table 5 reports the selection frequency for each perturbation.

Perturbation	Frequency	Selection Rate
Elastic*	984/1000	98.4%
Snow*	858/1000	85.8%
Glass Blur*	856/1000	85.6%
Motion Blur*	722/1000	72.2%
Color Shift*	442/1000	44.2%
Defocus Blur	287/1000	28.7%
Zoom Blur	284/1000	28.4%
Impulse Noise	222/1000	22.2%
Shot Noise	164/1000	16.4%
Glare	69/1000	6.9%
Gaussian Noise	47/1000	4.7%
Slight Rotation	22/1000	2.2%
Moiré	22/1000	2.2%
Frost	18/1000	1.8%
Fog	3/1000	0.3%

Table 5: Bootstrap stability analysis of perturbation selection (N=1000 resamples). * denotes the 5 perturbations selected on the full pilot sample. Selection rate indicates the fraction of bootstrap iterations in which each perturbation was included in the top-5 subset.

All five selected perturbations rank as the top-5 most frequently chosen across bootstrap iterations. Elastic deformation is the most stable (98.4%), followed by snow (85.8%) and glass blur (85.6%). Color shift has the lowest selection rate among the five (44.2%), reflecting its moderate MRD; however, it still substantially exceeds the highest non-selected perturbation (defocus blur at 28.7%). This gap confirms that color shift occupies a distinct niche not covered by higher-impact perturbations, validating its inclusion for category diversity. The clear separation between the selected and non-selected groups demonstrates that our perturbation selection is robust to sampling variability in the pilot data.

B Additional Figures

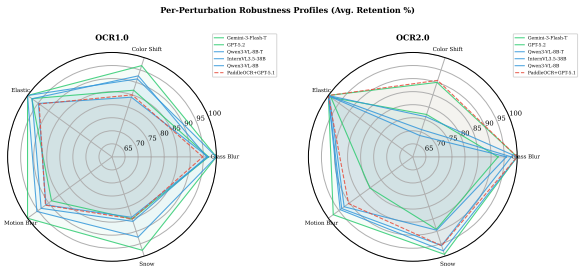

Figure 11: Per-perturbation robustness profiles showing average retention (%) across five perturbation types. Closed-source models (green) maintain near-circular profiles, while pipelines (red dashed) exhibit pronounced weaknesses.

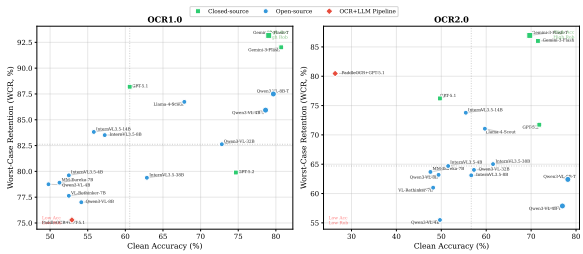

Figure 12: Clean accuracy vs. worst-case retention (WCR). Models in the upper-right quadrant achieve both high accuracy and strong robustness. Dashed lines indicate medians.