Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DocVLM-Fingerprint: Document VLM Faithfulness Under Visual Perturbations

Samarth Kukreja

samarthkukreja2003@gmail.com

Independent Researcher

Research Article

Keywords: Document VQA, Vision-Language Models, Faithfulness, Robustness, Visual Perturbations

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11757192/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

DocVLM-Fingerprint: Document VLM Faithfulness Under Visual Perturbations

Samarth Kukreja
Independent Researcher
Faridabad, India
samarthkukreja2003@gmail.com

Abstract

DocVLM-Fingerprint studies a narrow failure mode in document visual question answering (VQA): a vision-language model (VLM) can return a plausible short answer while its accompanying claim is unsupported by the annotated evidence. This work implements a reproducible evaluation pipeline for generated chart, optical character recognition (OCR) document, and scientific figure tasks. The pipeline includes four visual perturbations, open VLM clients, answer scoring, claim-support labels, plots, and failure records. The main run covers 120 generated examples, 480 perturbed variants, and three open models. Small clean subsets of ChartQA and DocumentVQA, together with a Qwen-only CharXiv perturbation supplement using real charts, provide external checks. The evidence supports a controlled diagnostic study of failure patterns, not a general ranking of VLMs.

Index Terms

document visual question answering, vision-language models, faithfulness, robustness, visual perturbations

I. Introduction

Most document visual question answering (VQA) evaluations report whether the final answer is correct. Answer accuracy alone can obscure two cases: a correct answer with an explanation unsupported by the annotated evidence, and an incorrect answer containing claims that are supported by that evidence. DocVLM-Fingerprint distinguishes these cases by reporting answer accuracy alongside claim support. We refer to the evaluated systems as vision-language models (VLMs).

The project is deliberately scoped as a systems and evaluation artifact. Its main output is the end-to-end evaluation workflow: task creation, perturbation generation, model evaluation, claim splitting, binary claim-support scoring, metric aggregation, plotting, and failure analysis. Deterministic mock outputs keep the pipeline testable on a laptop, while the Kaggle run supplies the open-model results used in the analysis.

The main contributions are the implemented evaluation pipeline and generated perturbation suite, with answer accuracy and claim support reported using bootstrap confidence intervals. The package also provides an audit worksheet for failure inspection, small external checks on ChartQA, DocumentVQA, and CharXiv, and descriptive analyses of disagreement between answer correctness and claim support.

The name *Fingerprint* refers to profiles of model weaknesses under perturbations and claim-support scoring. It does not refer to model ownership, identity verification, or watermark fingerprinting.

II. Related Work and Differentiation

Related work covers multimodal evaluation toolkits, hallucination benchmarks, robustness suites, optical character recognition (OCR) and chart tasks, and claim-based faithfulness measures. Examples include broad evaluation frameworks such as VLMEvalKit and lmms-eval [1], [2], hallucination benchmarks such as HallusionBench and POPE [3], [4], faithfulness measures such as FaithScore [5], robustness suites such as VLM-RobustBench and OCR-Robust [6], [7], and task benchmarks such as CharXiv, ChartQA, PlotQA, and DocVQA [8]–[11].

This work is closest to an integration study: it combines a small document VQA experiment, fixed perturbations, and the evaluation components described above in one reproducible package. The model comparisons exercise the pipeline; the central contribution is the reproducible diagnostic setup.

III. Dataset

The generated dataset contains 120 examples across three document VQA domains (Table I). Each record stores an image path, question, short answer, annotated evidence string, domain, and question type. The examples are deliberately controlled so that failures can be traced to the visual evidence without relying on ambiguous ground truth.

TABLE I
COMPOSITION OF THE GENERATED DATASET.

Domain	Examples
Charts/tables	40
OCR-heavy documents	40
Scientific figures/pages	40
Total	120

The chart subset contains bar charts, line charts, and table-style visuals. The OCR subset uses receipts, forms, invoice-style pages, and short policy pages. The scientific subset uses ablation tables, architecture diagrams, method-comparison figures, and method notes. The generated set avoids private PDFs and copyrighted paper screenshots.

IV. Visual Perturbations

For each example, the code creates four deterministic visual perturbations as input stress tests (Table II). These transformations are not treated as adversarial attacks or probes of model internals.

TABLE II
VISUAL PERTURBATIONS USED IN THE EVALUATION.

Perturbation	Purpose
Blur/downscale	Reduced resolution and softened details
Crop/removal	Partially missing visual evidence
JPEG compression	Blocking and quantization artifacts
Distractor text	Irrelevant overlaid text near evidence

The final generated run contains 480 perturbed variants, with 160 variants per domain. Metadata records the example ID, domain, perturbation type, severity, input path, and output path.

V. Claim Support Scoring

Answer accuracy can obscure claims unsupported by the annotated evidence, so the analysis also reports claim support.

The pipeline extracts claim text from the model-generated `Evidence:` component when available and decomposes that text into atomic claims. If this component is absent, it uses fallback parsing of the response. Each claim is stored as structured JavaScript Object Notation (JSON)-style data with claim text and a claim-type label. The current implementation uses only the factual and inference claim types.

Each claim receives one binary label: supported or unsupported. The scorer first applies exact, numeric, and entity checks, then compares the claim against the annotated evidence string. A cached large language model (LLM) judge is an optional fallback for inconclusive cases. No LLM-judge or cached-judge labels were used in the reported evaluation. The scorer evaluates support against the annotation; it does not independently verify claims against the underlying image. In the reported run, the scorer serves as an auditable component of the evaluation pipeline rather than a new faithfulness metric.

The reported metrics are answer accuracy, claim support, hallucination rate, and perturbation drop. Claim support is the fraction of scored claims labeled supported; hallucination rate is the fraction labeled unsupported. Here, the term hallucination rate denotes the unsupported-claim rate under this scorer, not an independent assessment of fabrication. Perturbation drop is the clean metric minus the perturbed metric.

VI. Experimental Setup

The archived model run evaluates Qwen3-VL 8B, InternVL3.5 8B (HF), and Qwen3-VL 4B on the generated examples and their four perturbed variants, for 600 cases per model and 1800 calls total. The corresponding repository identifiers are `qwen3_vl_8b`, `internvl35_8b_hf`, and `qwen3_vl_4b`, respectively. The evaluation includes two independent model families and a comparison between the Qwen3-VL 4B and 8B models. Table III lists the cases by domain using repository identifiers. Deterministic mock outputs support local validation without graphics processing unit (GPU) access. External data is reported separately.

TABLE III
GENERATED EVALUATION CASES BY DOMAIN.

Domain	Clean examples	Perturbed variants
chart	40	160
ocr_doc	40	160
scientific_figure	40	160

VII. Results

The Kaggle run evaluates all three models on the clean generated examples and all four perturbed variants. The parser recovered final answers for all 1800 calls; 1764 calls included explicit `Evidence:` lines, with most noncompliant generations occurring on heavily degraded chart inputs.

TABLE IV
MAIN METRICS FOR INTERNVL3.5 8B (HF) BY DOMAIN AND PERTURBATION.

Model	Domain	Perturbation	N	Acc.	Support	Halluc.
internvl35_8b_hf	chart	clean	40	0.70	1.00	0.00
internvl35_8b_hf	chart	blur_downscale	40	0.10	0.15	0.85
internvl35_8b_hf	chart	crop_removal	40	0.53	0.88	0.12
internvl35_8b_hf	chart	jpeg_compression	40	0.03	0.10	0.90
internvl35_8b_hf	chart	distractor_text	40	0.65	0.90	0.10
internvl35_8b_hf	ocr_doc	clean	40	0.95	1.00	0.00
internvl35_8b_hf	ocr_doc	blur_downscale	40	0.23	0.23	0.78
internvl35_8b_hf	ocr_doc	crop_removal	40	0.62	0.68	0.33
internvl35_8b_hf	ocr_doc	jpeg_compression	40	0.03	0.03	0.97
internvl35_8b_hf	ocr_doc	distractor_text	40	0.82	0.85	0.15
internvl35_8b_hf	scientific_figure	clean	40	0.88	1.00	0.00
internvl35_8b_hf	scientific_figure	blur_downscale	40	0.00	0.03	0.97
internvl35_8b_hf	scientific_figure	crop_removal	40	0.60	0.78	0.23
internvl35_8b_hf	scientific_figure	jpeg_compression	40	0.07	0.07	0.93
internvl35_8b_hf	scientific_figure	distractor_text	40	0.88	1.00	0.00

TABLE V
MAIN METRICS FOR QWEN3-VL 4B BY DOMAIN AND PERTURBATION.

Model	Domain	Perturbation	N	Acc.	Support	Halluc.
qwen3_vl_4b	chart	clean	40	0.95	0.97	0.03
qwen3_vl_4b	chart	blur_downscale	40	0.10	0.23	0.78
qwen3_vl_4b	chart	crop_removal	40	0.65	0.68	0.33
qwen3_vl_4b	chart	jpeg_compression	40	0.00	0.00	1.00
qwen3_vl_4b	chart	distractor_text	40	0.88	0.90	0.10
qwen3_vl_4b	ocr_doc	clean	40	0.90	0.90	0.10
qwen3_vl_4b	ocr_doc	blur_downscale	40	0.30	0.32	0.68
qwen3_vl_4b	ocr_doc	crop_removal	40	0.57	0.61	0.39
qwen3_vl_4b	ocr_doc	jpeg_compression	40	0.00	0.00	1.00
qwen3_vl_4b	ocr_doc	distractor_text	40	0.68	0.78	0.23
qwen3_vl_4b	scientific_figure	clean	40	0.93	1.00	0.00
qwen3_vl_4b	scientific_figure	blur_downscale	40	0.00	0.00	1.00
qwen3_vl_4b	scientific_figure	crop_removal	40	0.68	0.82	0.17
qwen3_vl_4b	scientific_figure	jpeg_compression	40	0.00	0.00	1.00
qwen3_vl_4b	scientific_figure	distractor_text	40	0.93	1.00	0.00

Figs. 1–4 summarize answer accuracy, claim support, the relationship between these metrics, and the drop in answer accuracy under perturbations, respectively.

A. Main Metrics

Tables IV–VI report the main metrics using the repository identifiers defined in Section VI. In Tables IV–VII, Support denotes claim support and Halluc. denotes the unsupported-claim rate defined above.

Bootstrap confidence intervals (CIs) use 2,000 case-level resamples with replacement within each perturbation (seed 20260703); the 95% percentile bounds are the linearly interpolated 2.5th and 97.5th percentiles of the resampled metric values. Table VII reports the aggregated intervals; the full table is stored in `results/metric_cis_by_perturbation.csv`.

TABLE VI
MAIN METRICS FOR QWEN3-VL 8B BY DOMAIN AND PERTURBATION.

Model	Domain	Perturbation	N	Acc.	Support	Halluc.
qwen3_vl_8b	chart	clean	40	1.00	1.00	0.00
qwen3_vl_8b	chart	blur_downscale	40	0.12	0.30	0.70
qwen3_vl_8b	chart	crop_removal	40	0.82	0.85	0.15
qwen3_vl_8b	chart	jpeg_compression	40	0.00	0.05	0.95
qwen3_vl_8b	chart	distractor_text	40	0.93	0.95	0.05
qwen3_vl_8b	ocr_doc	clean	40	0.97	1.00	0.00
qwen3_vl_8b	ocr_doc	blur_downscale	40	0.23	0.23	0.78
qwen3_vl_8b	ocr_doc	crop_removal	40	0.57	0.65	0.35
qwen3_vl_8b	ocr_doc	jpeg_compression	40	0.00	0.00	1.00
qwen3_vl_8b	ocr_doc	distractor_text	40	0.85	0.90	0.10
qwen3_vl_8b	scientific_figure	clean	40	0.97	0.97	0.03
qwen3_vl_8b	scientific_figure	blur_downscale	40	0.10	0.10	0.90
qwen3_vl_8b	scientific_figure	crop_removal	40	0.65	0.78	0.23
qwen3_vl_8b	scientific_figure	jpeg_compression	40	0.00	0.00	1.00
qwen3_vl_8b	scientific_figure	distractor_text	40	0.95	0.95	0.05

TABLE VII
BOOTSTRAP CONFIDENCE INTERVALS AGGREGATED BY PERTURBATION.

Perturbation	Acc. 95% CI	Support 95% CI	Halluc. 95% CI
clean	0.89 to 0.94	0.97 to 0.99	0.01 to 0.03
blur_downscale	0.09 to 0.17	0.14 to 0.22	0.78 to 0.86
crop_removal	0.58 to 0.68	0.70 to 0.79	0.21 to 0.30
jpeg_compression	0.00 to 0.03	0.01 to 0.05	0.95 to 0.99
distractor_text	0.80 to 0.88	0.88 to 0.94	0.06 to 0.12

B. Key Findings

The models answer most clean generated examples correctly. Average clean performance across the three domains is 0.842 answer accuracy and 1.000 claim support for InternVL3.5 8B (HF), 0.925 and 0.958 for Qwen3-VL 4B, and 0.983 and 0.992 for Qwen3-VL 8B, respectively.

JPEG compression yields the lowest performance, with 0.014 answer accuracy, 0.028 claim support, and 0.972 hallucination rate. Blur/downscale yields 0.131 answer accuracy, 0.175 claim support, and 0.825 hallucination rate. These values describe this run only; they should not be interpreted as guarantees about performance beyond the evaluated cases.

The other perturbations have smaller effects in this run: crop/removal averages 0.633 answer accuracy and 0.746 claim support, while distractor text averages 0.839 answer accuracy and 0.914 claim support.



Fig. 1. Answer accuracy by model, domain, and perturbation.

C. Clean Baseline

The clean rows in Tables IV–VI provide the baseline across charts, OCR documents, and scientific figures. Strong performance on these controlled inputs provides context for the declines under specific visual degradations.

D. Perturbation Robustness

The perturbed rows compare performance in each domain with its clean baseline under the same four transformations. JPEG compression and blur/downscale produce the largest drops. Crop/removal produces moderate declines, while distractor text is less disruptive in this generated set. These differences are observations from this run, not causal estimates about model internals.



Fig. 2. Claim support by model, domain, and perturbation.

E. Accuracy vs. Claim Support

The scatter plot shows the relationship between answer accuracy and claim support. The two-line answer-and-evidence protocol allows the final answer to be scored separately from the extracted evidence claim. In this run, the clearest disagreement consists of 113 claim records for which the final answer is incorrect while the extracted evidence claim is supported by the annotated evidence. The main disagreement is therefore not correct answers accompanied by unsupported evidence. These records show that an extracted evidence claim can be supported by the annotation even when the final answer is incorrect, a distinction that answer accuracy alone would obscure.

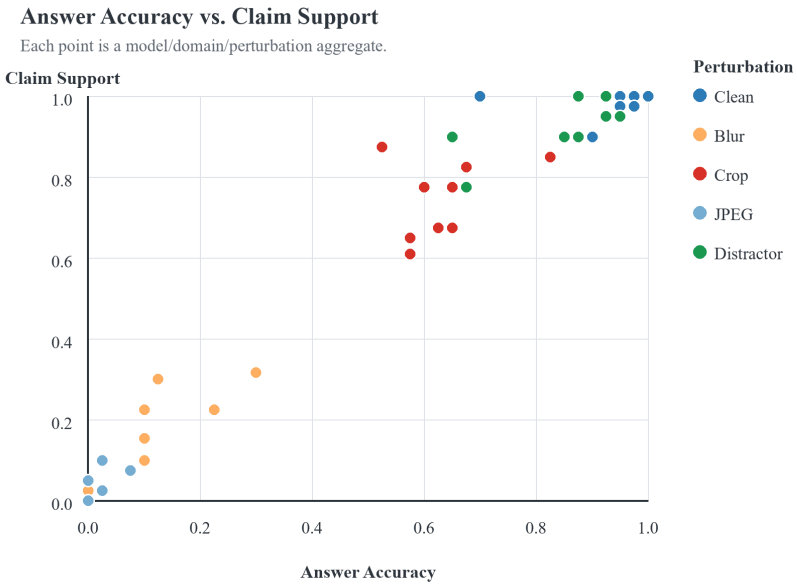


Fig. 3. Relationship between answer accuracy and claim support.

F. Qwen Size Comparison

Within the Qwen3-VL family, average clean answer accuracy is 0.925 for the 4B model and 0.983 for the 8B model; average clean claim support is 0.958 and 0.992, respectively. This comparison examines sensitivity to model size within one family; it does not represent a third independent model family.

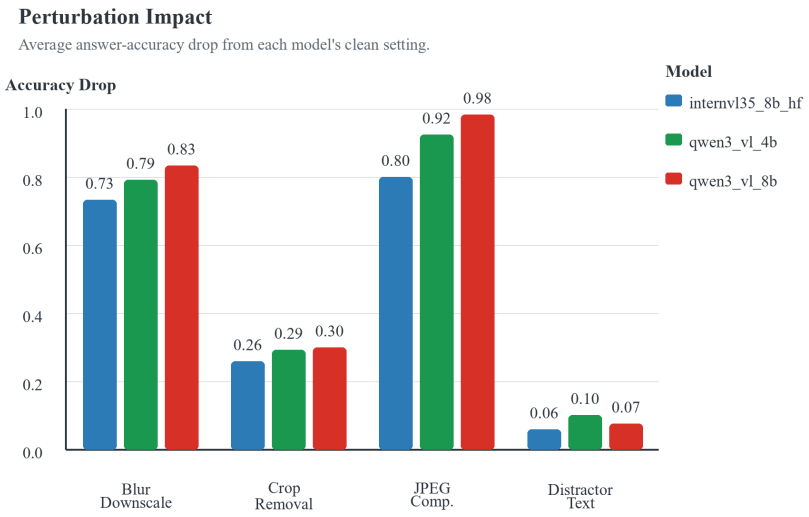


Fig. 4. Average drop in answer accuracy from the clean setting, grouped by perturbation and model.

G. External Real-Data Checks

The external checks are reported separately from the generated set. A 30-example `HuggingFaceM4/ChartQA` subset evaluates clean chart question answering for all three models, giving 90 additional VLM calls. Exact-match answer accuracies are 0.600 for Qwen3-VL 8B, 0.200 for InternVL3.5 8B (HF), and 0.500 for Qwen3-VL 4B. Results are stored in `results/external/chartqa`.

Here, `HuggingFaceM4/DocumentVQA` denotes a preprocessed Hugging Face distribution of DocVQA. A 30-example validation subset evaluates clean OCR document question answering for the same three models, giving 90 additional VLM calls. Answer accuracies are 0.667 for Qwen3-VL 8B, 0.633 for InternVL3.5 8B (HF), and 0.633 for Qwen3-VL 4B. Results are stored in `results/external/documentvqa`.

A separate CharXiv supplement contains 20 real arXiv chart examples with the same four perturbations, evaluated for Qwen3-VL 4B and Qwen3-VL 8B only. Clean CharXiv accuracy is 0.400 and 0.450, respectively; JPEG compression reduces both to 0.100, while accuracy under blur/downscale is 0.150 and 0.250. Results are stored in `results/external/charxiv`.

CharXiv provides the most direct external check of transfer to perturbed real charts, whereas the ChartQA and DocumentVQA checks are clean-only. CharXiv images remain local-only and are excluded from the public release. These small external checks are not merged with the generated perturbation study.

H. Descriptive Analysis

The analysis helper writes descriptive tables to `results/ablations`. These post-hoc summaries analyze existing evaluation records without rerunning the VLMs under modified experimental conditions. The tables summarize counts of disagreements between answer correctness and claim support, scorer-method usage by label, perturbation effect sizes, and cases with no extracted claims.

They aid interpretation of the run and are not proposed as a new metric.

I. Failure Examples

1) **Chart example** (`chart_002`; `distractor_text`).
   Question: Which category has the highest value?
   Model response:
   > Answer: NYB
   > Evidence: The bar labeled "NYB" is the tallest, with a value of 22, which is higher than the other categories.

   Reference answer: HYB.
   Extracted claim:
   > The bar labeled "NYB" is the tallest, with a value of 22, which is higher than the other categories.

   Explanation: The response selects the distractor label NYB rather than HYB. The binary scorer labels the extracted claim unsupported by the annotated evidence.
2) **Receipt example** (`ocr_001`; `distractor_text`).
   Question: What is the receipt total?
   Model response:
   > Answer: 52
   > Evidence: The receipt shows "TOTAL" with "52" highlighted in red, indicating it is the correct total.

   Reference answer: 50.
   Extracted claims:
   > The receipt shows "TOTAL" with "52" highlighted in red, indicating it is the correct total.
   > The receipt shows "TOTAL" with the amount "52" highlighted in red.

   Explanation: The response reports 52 rather than the reference total of 50. The binary scorer identifies unsupported claim text in this response when evaluated against the annotated evidence.
3) **Scientific figure example** (`sci_008`; `distractor_text`).
   Question: Which method has the highest accuracy?
   Model response:
   > Answer: BASE
   > Evidence: The BASE method shows the highest accuracy value of 76 in the table.

   Reference answer: FULL.
   Extracted claim:
   > The BASE method shows the highest accuracy value of 76 in the table.

   Explanation: The response selects BASE rather than FULL. The binary scorer labels the extracted claim unsupported by the annotated evidence.

The generated plots are stored in `results/figures` and copied into `report/figures` for paper builds.

VIII. Failure Analysis

The examples in Section I illustrate claims labeled unsupported by the binary scorer under distractor-text perturbations. These are debugging examples, not a taxonomy of VLM failures.

The manual audit worksheet is prepared for post-run review; its findings should be summarized

only after the records have been manually checked.

IX. Limitations

The study is intentionally small. The main run covers all three generated domains for 3 open models, but only two independent model families are represented because Qwen3-VL 4B and 8B form a size comparison. The generated data improves control and reproducibility but limits what can be concluded about real documents. Most questions are single-hop short-answer lookups. The relationship between answer accuracy and claim support may differ for longer rationales; that setting was not evaluated here.

Some generations failed to follow the requested *Evidence:* line format, which is reported as a protocol-compliance limitation. The binary scorer is transparent and testable, although it can miss paraphrases or ambiguous visual evidence. An unsupported label means that the scorer does not find the claim entailed by the annotated evidence string. The reported hallucination rate therefore measures unsupported claims under this scorer; it does not independently establish whether those claims are fabricated.

The ChartQA, DocumentVQA, and CharXiv subsets provide real-data checks, but they are small and separate from the generated perturbation study; CharXiv is also Qwen-only. The simulated JPEG perturbation can remove small text entirely, so near-zero scores should be interpreted as a missing-information floor condition rather than calibrated compression robustness.

Future work should add independent multi-annotator review and further visual transformations, including rotations, layout shifts, lighting variation, watermarks, and adversarial text.

X. Code and Data Availability

The project repository is available at https://github.com/SamarthKukreja/DocVLM-Fingerprint. The archived release is available at https://doi.org/10.5281/zenodo.21433621. The generated-dataset results in this report correspond to repository release v1.0.0. The main generated dataset, perturbation images, model outputs, metrics, confidence intervals, figures, ablation tables, report source, and tests are included in the repository so the reported artifacts can be inspected without rerunning VLM inference.

External ChartQA, DocumentVQA, and CharXiv checks are reported separately from the main generated-dataset run. The repository does not redistribute third-party-derived external dataset images unless licensing permits it; those image folders are treated as local-only artifacts. The external result files and metadata are included only where they do not require redistribution of third-party image assets.

XI. Ethics and Licensing

The main dataset is generated by the project and is intended for controlled reliability analysis, not for measuring human ability or ranking deployed systems. The code is released under the MIT license. External datasets retain their original licenses and attribution requirements. Because ChartQA, DocumentVQA, and CharXiv include third-party data sources, their images should be downloaded or regenerated by users from authorized sources and should not be redistributed from

this repository without a separate license review.

The evaluation can identify claims unsupported by the annotated evidence. The binary scorer assigns support labels using the annotated evidence string and should not be used as a safety certification tool. Reported hallucination rates denote unsupported-claim rates under the project scorer, not definitive assessments of real-world model behavior.

XII. Conclusion

DocVLM-Fingerprint provides a reproducible workflow for studying document VLM behavior under visual perturbations. It combines 120 generated examples and four perturbations per example with model client interfaces, claim-support scoring, metrics, plots, failure examples, and a small inspection dashboard. The package can be rerun, audited, and extended with other models or datasets.

References

[1] OpenCompass Contributors, "VLMEvalKit: An open-source toolkit for evaluating large multi-modality models," 2024, project repository: github.com/open-compass/VLMEvalKit.

[2] EvolvingLMMs-Lab Contributors, "lmms-eval: Evaluation suite for large multimodal models," 2024, project repository: github.com/EvolvingLMMs-Lab/lmms-eval.

[3] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou, "HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," 2023, arXiv:2310.14566.

[4] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of EMNLP*, 2023.

[5] L. Jing, R. Li, Y. Chen, M. Jia, and X. Du, "FaithScore: Evaluating hallucinations in large vision-language models," 2023, arXiv:2311.01477.

[6] "VLM-RobustBench: Robustness evaluation for vision-language models," 2026, arXiv:2603.06148.

[7] "OCR-Robust: Robustness evaluation for ocr-centric vision-language tasks," 2026, arXiv:2606.26041.

[8] "CharXiv: Chart and figure understanding for vision-language models," 2024, arXiv:2406.18521.

[9] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the Association for Computational Linguistics: ACL*, 2022.

[10] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "PlotQA: Reasoning over scientific plots," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020.

[11] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for vqa on document images," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021.