



WildDoc: How Far Are We from Achieving Comprehensive and Robust Document Understanding in the Wild?

An-Lan Wang^{1,†}, Jingqun Tang^{1,†,✉}, Lei Liao^{1,†}, Hao Feng¹, Qi Liu¹, Xiang Fei¹,
Jinghui Lu¹, Han Wang¹, Hao Liu¹, Yuliang Liu², Xiang Bai², Can Huang¹

¹ByteDance, China

²Huazhong University of Science and Technology, China

{wanganlan, tangjingqun, can.huang}@bytedance.com, {ylliu, xbai}@hust.edu.cn

Abstract

The rapid advancements in Multimodal Large Language Models (MLLMs) have significantly enhanced capabilities in Document Understanding. However, prevailing benchmarks like DocVQA and ChartQA predominantly comprise *scanned or digital* documents, inadequately reflecting the intricate challenges posed by diverse real-world scenarios, such as variable illumination and physical distortions. This paper introduces WildDoc, the inaugural benchmark designed specifically for assessing document understanding in natural environments. WildDoc incorporates a diverse set of manually captured document images reflecting real-world conditions and leverages document sources from established benchmarks to facilitate comprehensive comparisons with digital or scanned documents. Further, to rigorously evaluate model robustness, each document is captured four times under different conditions. Evaluations of state-of-the-art MLLMs on WildDoc expose substantial performance declines and underscore the models' inadequate robustness compared to traditional benchmarks, highlighting the unique challenges posed by real-world document understanding. Our project homepage is available at <https://bytedance.github.io/WildDoc>.

1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly enhanced their capabilities across various vision-language tasks. Notably, recent studies [8, 10, 24, 46, 50, 51, 53] have extended the application of MLLMs from processing basic low-resolution images to comprehending high-resolution document images [8, 10, 22, 24, 46], marking a significant shift in their scope of applicability.

Despite these technological strides, prevalent benchmarks for document understanding [12], e.g.,

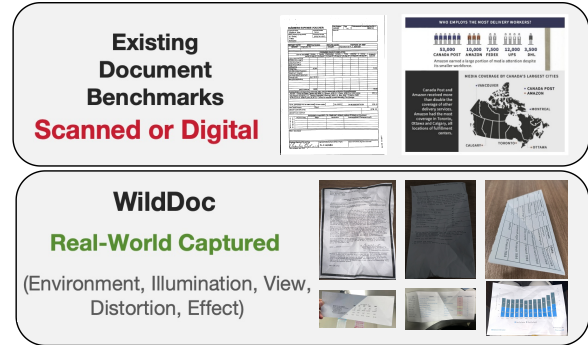


Figure 1: Comparison of WildDoc with existing benchmarks for document understanding, highlighting the predominance of scanned or digital document images in current benchmarks versus the real-world captured document images in WildDoc.

DocVQA [30], InfoVQA [29], ChartQA [28], are predominantly composed of *scanned or digital* documents (see Figure 1, top). These benchmarks fail to capture the challenges posed by documents in the real world, which often involves photo capturing of paper documents and screen capturing of electronic records, each introducing complexities such as variable views, illumination, and physical distortions. Consequently, these limitations prompt critical inquiries regarding the efficacy of current models under real-world conditions, leading us to question: *How far are we from achieving comprehensive and robust document understanding in the wild?*

To address this question, we introduce WildDoc, the first benchmark focusing on document understanding in the real world, as depicted in Figure 1 (bottom). This benchmark boasts a meticulously curated collection of over 12,000 document images that reflect a broad spectrum of real-world scenarios. These real-world photographic factors are mainly categorized into five: Environment, Illumination, View, Distortion, and Effect, each with multiple variations to thoroughly simulate real-world

[†]equal contribution. [✉] corresponding author.

complexities (detailed in Table 1).

Moreover, WildDoc utilizes the same document sources as three widely used benchmarks [28–30], which offer three advantages: 1) It can cover a variety of common document types, i.e., regular documents, charts, and tables; 2) It allows for the reuse of existing question-answer pairs from these benchmarks, thereby reducing annotation efforts; 3) It facilitates direct and fair comparisons between scanned/digital and real-world document understanding capabilities, thereby highlighting performance discrepancies. Additionally, we introduce a consistency metric designed to evaluate the robustness of model performance across varied real-world conditions. Specifically, each document is captured under four distinct scenarios, and this metric measures its ability to consistently provide accurate answers.

Based on WildDoc, we conduct experiments to evaluate numerous representative MLLMs, including general MLLMs (e.g., Qwen2.5-VL [32]) and the leading closed-course MLLMs (e.g., GPT-4o [31], Doubao-1.5-pro [15]). The experiment results demonstrate that (1) Existing MLLMs exhibit a large performance decline in WildDoc compared to traditional document benchmarks, with models like GPT-4o showing an average performance decrease of 35.3. (2) Existing MLLMs demonstrate inadequate robustness in document understanding. This is evident from their lower scores in consistency evaluations, with Doubao-1.5-pro achieving the highest score of 50.6. (3) Some models exhibit minimal performance variations and tend to saturate on the original benchmark, yet they experience significant performance declines and disparities on WildDoc. As a result, these findings reveal that there is still a large room for comprehensive and robust document understanding in the wild, and highlight the value of WildDoc.

Our contributions are summarized as follows:

- We establish WildDoc, a benchmark designed to systematically evaluate the document understanding ability of existing MLLMs, which provides the community with fresh insights on document understanding in the real world.
- To thoroughly evaluate existing models, we further propose a new robustness metric – Consistency Score. This metric evaluates whether the model can consistently handle varying real-world situations.

Factor	Choices
Environment	Indoor, Outdoor
Illumination	Light, Dark Flashlight On, Flashlight Off
View	Top, Down, Left, Right, etc.
Distortion	Crease, Wrinkle, Bend, etc.
Effect	Shadows, Overexposure, Blur, etc.

Table 1: The five most common factors affecting document understanding in real-world scenarios. For each factor, various choices are further provided to illustrate the range of possible conditions.

- We benchmark numerous advanced MLLMs on WildDoc, revealing significant potential for improvement in robust document understanding.

2 Related Works

2.1 MLLMs for Document Understanding

Multimodal Large Language Models [2, 6, 9–11, 13, 21, 24–26, 32, 35, 40–43] have demonstrated remarkable performance across a range of vision-language tasks, particularly distinguished by their exceptional zero-shot capabilities. Beyond these tasks, the problem of understanding documents has received a fair amount of interest recently [22]. For example, early works like LLaVAR [50] extend LLaVA [23] into the realm of document understanding by tuning in collected document datasets. Furthermore, DocPedia [10] introduces higher input resolution by leveraging frequency information, achieving remarkable performance. More recently, mPLUG-DocOwl [46], TextMonkey [24], IXC4KHD [8], TextSquare [35], and Vary [44] further enhance the document understanding ability by leveraging large-scale document-related datasets and increasing the input resolution. TextHarmony [53] further unifies the perception, understanding, and generation of visual text. Despite the promising results achieved by the above-mentioned MLLMs in the document understanding area, their document understanding capabilities in real-world scenarios are not fully validated. This is primarily due to the lack of benchmarks for documents in the wild.

2.2 Document Understanding Benchmarks

Existing document understanding benchmarks [12, 18, 30, 33, 36] can be mainly divided into two cate-

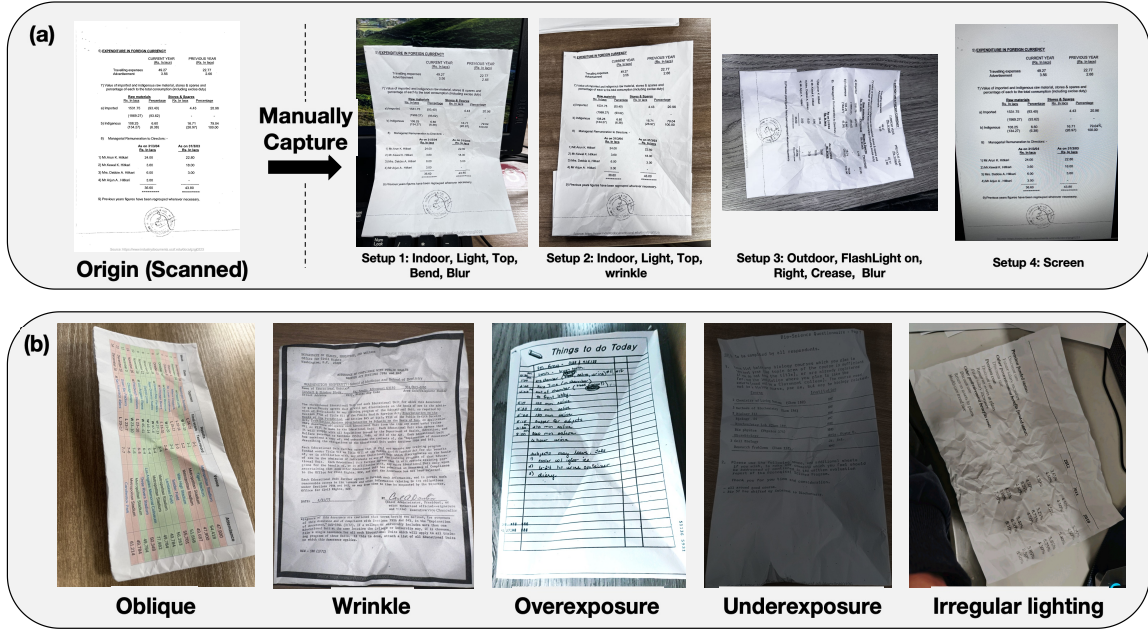


Figure 2: Overview of the WildDoc. (a) For every document, we manually capture four images under different setups. (b) Several representative examples that encompass different real-world conditions. More examples are listed in the Appendix.



Figure 3: Statistics on image capture setup.

gories according to the type of images: (1) scanned document images [30], which contains images that are scanned and binarized. (2) digital document images [17, 18, 29]. For example, TableVQA-Bench [18] uses a rendering tool to collect synthetic images in the Wikipedia style, and AI2D [17] crawls images from Google Image Search. Despite playing a crucial role in benchmarking document understanding, these benchmarks overlook the gap between scanned/digital documents and real-world captured documents, thus, they are unable to accurately assess the performance of current models on documents encountered in the real world. In

this work, we establish WildDoc, the first document benchmarks that focus on real-world captured document images. It contains manually captured document images from different scenarios.

3 WildDoc Benchmark

In this section, we detail the data collection and filtering process and present some statistics.

3.1 Data Collection

Firstly, we introduce the raw document source of WildDoc, which aims to ensure a broad coverage of the various types of documents encountered in everyday life. Specifically, our focus is primarily on documents from three previous benchmarks, DocVQA [30], ChartQA [28], and TableVQA [18]. Utilizing these documents offers two main benefits: 1) Reusing the Question-Answer pairs from these benchmarks reduces the burden of annotation, and 2) It allows for a direct and fair comparison between WildDoc and these benchmarks, thereby highlighting the performance gap.

Next, we detail the document image capture process. Prior to image capture, all documents are printed using high-resolution printers to preserve the original text clarity and layout nuances, and each document is carefully trimmed, adjusting its physical dimensions. To ensure the captured images in our benchmark cover a wide range of sce-

narios encountered in daily life, we selected five key factors (i.e., *environment*, *illumination*, *view*, *distortion*, *effect*) that are common in the real world, and offers multiple choices for each, as listed in Table 1. Figure 2 provides several examples. During the image capture sessions, each participant adheres to predefined but varied setups. Additionally, we do not restrict the types of image capture equipment used in the data collection process; instead, we embrace the diversity of equipment, which allows us to collect data that better represents the varying qualities of images, ultimately enhancing the diversity of WildDoc.

3.2 Data Filtering

Upon completing all data collection processes, we convened a special panel of quality inspectors to review all collected images to ensure their effectiveness. Before the formal review, all quality inspectors are provided with a detailed explanation of the review guidelines. Essentially, every captured document image must meet two fundamental requirements: 1) Adherence to the specified setup, and 2) The image content must allow for the corresponding annotated questions to be answered accurately. Images that fail to meet these criteria are returned for recapture. Based on these criteria, the data filtering stage consists of two rounds:

In the *first round* of reviews, the primary focus is to rigorously evaluate whether the captured documents adhere to the specified setup and ensure that no parts of the documents are missing. This baseline check does not require extensive expertise from the inspectors, as it mainly revolves around adherence to explicit, predefined procedural standards rather than subjective interpretations.

In the *second round* of reviews, quality inspectors are tasked with a thorough assessment of the captured documents in conjunction with the corresponding question-answer pairs, to ensure that the answers can indeed be accurately derived from the document images. This rigorous scrutiny validates the relevance and accuracy of the benchmark. In this round, each inspector must be proficient in English to effectively comprehend the documents, questions, and answers.

3.3 Data Quality

OCR accuracy. We quantify OCR [34, 37–39, 52] accuracy on 100 randomly selected documents from DocVQA and WildDoc benchmarks using

PaddleOCR [4], noting a 20.2% decline in Line-level accuracy in WildDoc.

Image quality. We employ LIQE [49], a no-reference quality metric correlating strongly with human perception, to rate 1000 images on a scale from 0 (Bad) to 4 (Perfect). The average quality score of DocVQA is 3.40, compared to 1.57 for WildDoc.

Answerability. To verify the consistency between WildDoc and the original dataset, a human performance validation on the DocVQA subset (1000 randomly selected questions) is performed. Participants achieve 97.2% accuracy on WildDoc versus 98.1% on original scans. The marginal 0.9% gap confirms that performance drops in MLLMs stem from understanding limitations rather than irrecoverable information loss.

3.4 Data Statistics

In Figure 2, we provide an overview of WildDoc. The construct benchmark comprises over 12000 document images. In Figure 3 (a), the distribution of the image capture setups is visualized, where we maintain a diverse and balanced distribution of choices for different factors, which enhances the reliability of WildDoc.

More statistics are illustrated in the Appendix A.

4 Experiments

4.1 Metrics

Accuracy and ANLS. Following previous benchmarks [18, 28, 30], we report the Average Normalized Levenshtein Similarity (ANLS) and Accuracy (Acc.).

Consistency score. In WildDoc, we manually capture four images for each document with different setups. This enables us to evaluate the robustness of models when handling different real-world scenarios. Specifically, for one question, the model must correctly answer the question based on each of the images for its response to be considered positive; otherwise, it is considered negative. The consistency score offers a more precise evaluation of the model’s performance, reflecting its capability in robust document understanding.

More details are provided in the appendix A.

4.2 Main Results

Table 2 presents the performance of several state-of-the-art open-source and closed-source MLLMs. All models suffer a decline in all three subsets,

Model	DocVQA			ChartQA			TableVQA			AVG.	
	Origin	WildDoc		Origin	WildDoc		Origin	WildDoc		Acc.	Consistency
	ANLS	ANLS	Consistency	Acc.	Acc.	Consistency	Acc.	Acc.	Consistency		
MiniMonkey-2B [2024]	86.5	54.3 -32.2	22.8 -31.5	73.5	32.3 -41.2	12.0 -20.3	51.1	31.3 -19.8	13.4 -17.9	39.3	16.1 -23.2
Monkey [2024]	56.6	31.0 -25.6	9.9 -21.1	55.7	22.4 -33.3	9.8 -12.6	33.4	23.0 -10.4	11.7 -11.3	25.4	10.5 -14.9
Phi-3.5-Vision [2024]	70.4	30.7 -39.7	11.9 -18.8	71.5	29.1 -42.4	12.4 -16.7	59.7	28.6 -31.1	11.1 -17.5	29.5	11.8 -17.7
TextHarmony [2024]	49.2	37.1 -12.1	16.0 -11.1	38.6	21.9 -16.7	10.2 -11.7	20.1	14.7 -5.4	6.5 -8.2	24.6	10.9 -13.7
mPLUG-Owl3 [2024]	46.2	27.7 -18.5	11.2 -16.5	40.2	22.8 -17.4	11.0 -11.8	21.5	16.8 -4.7	7.9 -8.8	22.4	10.0 -12.4
Janus-Pro-7B [2025]	40.9	19.5 -21.4	5.8 -13.7	25.1	12.3 -12.8	6.6 -5.7	33.1	20.2 -12.9	13.8 -6.4	17.4	8.7 -8.7
Llava-Onevision-7B [2024]	87.2	52.9 -34.3	23.7 -29.2	80.3	49.4 -30.9	20.2 -29.2	62.7	33.9 -28.8	13.3 -20.6	45.4	19.1 -26.3
GLM-4V-9B [2024]	81.0	66.5 -14.5	50.3 -16.2	30.1	23.0 -7.1	14.8 -8.2	61.1	46.5 -14.6	28.4 -18.1	45.4	31.2 -14.2
MiniCPM-V2.6 [2024]	90.1	62.9 -27.2	32.3 -30.6	79.5	43.6 -35.9	19.1 -24.5	68.3	43.5 -24.8	19.2 -24.3	50.0	23.5 -26.5
SAIL-VL-2B [2025]	86.1	49.8 -36.3	21.1 -28.6	80.3	49.4 -30.9	14.7 -34.7	64.6	35.0 -29.6	15.3 -19.7	44.7	17.0 -27.7
InternLM-XC2.5 [2024]	90.4	54.3 -36.1	32.3 -22.0	81.9	40.6 -41.3	19.1 -21.5	71.8	38.8 -33.0	14.9 -23.9	44.6	22.1 -22.5
Ovis1.6-Gemma2-9B [2024]	88.9	58.0 -30.9	28.4 -29.6	81.1	49.4 -31.7	18.6 -30.8	47.2	28.2 -19.0	12.5 -15.7	45.2	19.8 -25.4
InternVL2.5-8B-MPO [2024]	92.1	59.6 -32.5	27.6 -32.0	83.1	41.6 -41.5	18.6 -23.0	70.1	38.6 -31.5	12.1 -26.5	46.6	19.4 -27.2
Qwen2.5-VL-7B [2025]	93.9	79.8 -14.1	62.6 -17.2	87.6	64.8 -22.8	39.0 -25.8	75.9	57.2 -18.7	35.2 -22.0	67.3	45.6 -21.7
InternVL2.5-78B-MPO [2024]	95.4	69.5 -25.9	42.8 -26.7	88.3	43.8 -44.5	28.9 -14.9	76.8	45.8 -31.0	19.8 -26.0	53.0	30.5 -22.5
Qwen2.5-VL-72B [2025]	<u>95.5</u>	<u>80.3</u> -15.2	<u>63.1</u> -17.2	<u>89.5</u>	<u>66.5</u> -23.0	<u>45.5</u> -21.0	83.2	<u>64.8</u> -18.4	<u>40.4</u> -24.4	<u>70.6</u>	<u>49.7</u> -20.9
Closed-source MLLMs											
GPT-4o [2024]	91.5	63.2 -28.3	39.5 -23.7	86.7	30.3 -56.4	20.6 -9.7	75.7	54.4 -21.3	27.0 -27.4	49.3	29.0 -20.3
Gemini-1.5-pro [2024]	92.4	81.0 -11.4	68.6 -12.4	81.3	30.6 -50.7	20.8 -9.8	80.0	67.3 -12.7	46.2 -21.1	59.6	45.2 -14.4
Claude3.5 sonnet [2024]	95.4	54.3 -41.1	25.9 -28.4	90.8	37.1 -53.7	24.1 -13.0	82.1	45.1 -37.0	26.5 -18.6	45.5	25.5 -20.0
Doubao-1.5-pro [2025]	96.9	77.3 -19.6	57.3 -20.0	89.1	79.5 -9.6	66.6 -12.9	<u>82.6</u>	64.6 -18.0	<u>41.2</u> -23.4	73.7	55.0 -18.7

Table 2: Performance of the leading MLLMs. We report the results on WildDoc and the corresponding results in the original benchmark. The details of the consistency score are illustrated in the metrics section. “AVG.” indicates the average results on WildDoc. The top result is **bolded**, while the second-best is underlined.

GPT-4o suffers a decline of 28.3, 56.4, 21.3 in the three subsets, respectively. The results indicate that current MLLMs have not yet achieved satisfactory levels of document understanding capability when handling real-world scenarios. Among these models, Doubao-1.5-pro [15] stands out with an average accuracy of 73.7%, and Qwen2.5-VL-72B achieves the second highest average accuracy.

Additionally, we have an interesting finding that some models exhibit similar performance on the original dataset, yet display significant discrepancies when evaluated on WildDoc. For example, both InternVL2.5-78B-MPO and Claude3.5 sonnet score 95.4 on the original DocVQA benchmark, yet this difference expands to 15.2 points on the WildDoc benchmark. Furthermore, we select the top five models based on their performance on the WildDoc and calculate their mean and standard deviation on the original DocVQA benchmark, which are 94.98 and 0.612, respectively. This suggests that DocVQA may offer limited insights into the performance differences among the models. In contrast, on WildDoc, these values are 76.0 and 6.1, indicating a broader dispersion and more distinct performance variations among the models. These results further highlight the value of WildDoc in benchmarking the document understanding ability.

For the robustness evaluation, all models suffer a further decline. Notably, Doubao-1.5-pro records the highest average consistency score of 55.0, indicating a large room for improvement for current

MLLMs.

5 Conclusion

To thoroughly evaluate the performance of existing models, in this work, we establish the first real-world document understanding benchmark, WildDoc, which incorporates over 12K manually captured document images that cover various real-world factors. Based on WildDoc, we evaluate several state-of-the-art MLLMs. The results show that there is a large performance gap between scanned/digital and real-world document understanding, suggesting substantial opportunities for enhancement. We aspire that WildDoc will offer the research community fresh insights.

Limitations

The document source of WildDoc is derived from three widely used document benchmarks, which may hinder the coverage of real-world documents. Additionally, it’s important to note that our study is concentrated solely on English, which may limit the broader application of our benchmark and findings to other languages.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat

- Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [3] Anthropic. 2024. **Claude 3.5 sonnet**. Accessed: 2025-02-13.
- [4] PaddlePaddle Authors. 2023. Paddleocr: A versatile ocr toolkit with 80+ languages recognition. <https://github.com/PaddlePaddle/PaddleOCR>.
- [5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- [6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- [7] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. 2025. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*.
- [8] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.
- [9] Xiang Fei, Jinghui Lu, Qi Sun, Hao Feng, Yanjie Wang, Wei Shi, An-Lan Wang, Jingqun Tang, and Can Huang. 2025. Advancing sequential numerical prediction in autoregressive models. *arXiv preprint arXiv:2505.13077*.
- [10] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*.
- [11] Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, et al. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*.
- [12] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, et al. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.
- [13] Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. 2025. Vispeak: Visual instruction feedback in streaming videos. *arXiv preprint arXiv:2503.12769*.
- [14] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- [15] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. 2025. Seed1. 5-v1 technical report. *arXiv preprint arXiv:2505.07062*.
- [16] Mingxin Huang, Yuliang Liu, Dingkan Liang, Lianwen Jin, and Xiang Bai. 2024. Minimonkey: Alleviating the semantic sawtooth effect for lightweight mllms via complementary image pyramid. *arXiv preprint arXiv:2408.02034*.
- [17] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*.
- [18] 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*.
- [19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- [20] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773.
- [21] Kun-Yu Lin, Hongjun Wang, Weining Ren, and Kai Han. 2025. Panoptic captioning: Seeking an equivalency bridge for image and text. *arXiv preprint arXiv:2505.16334*.
- [22] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. 2024. Hrvda: High-resolution visual document assistant. In *CVPR*.
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *NerulIPS*.
- [24] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

- [25] Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, et al. 2024. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*.
- [26] Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, et al. 2025. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mlm reasoning. *arXiv preprint arXiv:2505.15154*.
- [27] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.
- [28] 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. *arXiv preprint arXiv:2203.10244*.
- [29] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *WACV*.
- [30] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.
- [31] OpenAI. 2024. [Gpt-4o main page](#). Accessed: 2025-02-13.
- [32] Team Qwen. 2025. [Qwen2.5-vl main page](#). Accessed: 2025-02-13.
- [33] Bin Shan, Xiang Fei, Wei Shi, An-Lan Wang, Guozhi Tang, Lei Liao, Jingqun Tang, Xiang Bai, and Can Huang. 2024. Mctbench: Multimodal cognition towards text-rich visual scenes benchmark. *arXiv preprint arXiv:2410.11538*.
- [34] Jingqun Tang, Weidong Du, Bin Wang, Wenyang Zhou, Shuqi Mei, Tao Xue, Xing Xu, and Hai Zhang. 2023. Character recognition competition for street view shop signs. *National Science Review*, 10(6):nwad141.
- [35] Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, et al. 2024. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*.
- [36] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.
- [37] Jingqun Tang, Wenming Qian, Luchuan Song, Xiena Dong, Lan Li, and Xiang Bai. 2022. Optimal boxes: boosting end-to-end scene text recognition by adjusting annotated bounding boxes via reinforcement learning. In *European Conference on Computer Vision*, pages 233–248. Springer.
- [38] Jingqun Tang, Su Qiao, Benlei Cui, Yuhang Ma, Sheng Zhang, and Dimitrios Kanoulas. 2022. [You can even annotate text with voice: Transcription-only-supervised text spotting](#). In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 4154–4163, New York, NY, USA. Association for Computing Machinery.
- [39] Jingqun Tang, Wenqing Zhang, Hongye Liu, Mingkun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. 2022. Few could be better than all: Feature sampling and grouping for scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4572.
- [40] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- [41] An-Lan Wang, Bin Shan, Wei Shi, Kun-Yu Lin, Xiang Fei, Guozhi Tang, Lei Liao, Jingqun Tang, Can Huang, and Wei-Shi Zheng. 2024. Pargo: Bridging vision-language with partial and global views. *arXiv preprint arXiv:2408.12928*.
- [42] Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. 2025. Vision as lora. *arXiv preprint arXiv:2503.20680*.
- [43] Junqiao Wang, Zeng Zhang, Yangfan He, Yuyang Song, Tianyu Shi, Yuchen Li, Hengyuan Xu, Kunyu Wu, Guangwu Qian, Qiuwu Chen, et al. 2024. Enhancing code llms with reinforcement learning in code generation. *arXiv preprint arXiv:2412.20367*.
- [44] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2025. Vary: Scaling up the vision vocabulary for large vision-language model. In *ECCV*.
- [45] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- [46] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

- [47] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingen Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.
- [48] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.
- [49] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081.
- [50] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.
- [51] Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, et al. 2025. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37:7185–7212.
- [52] Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Can Huang, Hao Liu, Xin Tan, Zhizhong Zhang, and Yuan Xie. 2024. Multi-modal in-context learning makes an ego-evolving scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15567–15576.
- [53] Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Zhizhong Zhang, Can Huang, and Yuan Xie. 2024. Harmonizing visual text comprehension and generation. *arXiv preprint arXiv:2407.16364*.

A Appendix

A.1 More Statistics

We present statistics regarding the image capture equipment used. As illustrated in Figure 4, we ensure a diverse range of image capture devices are maintained.

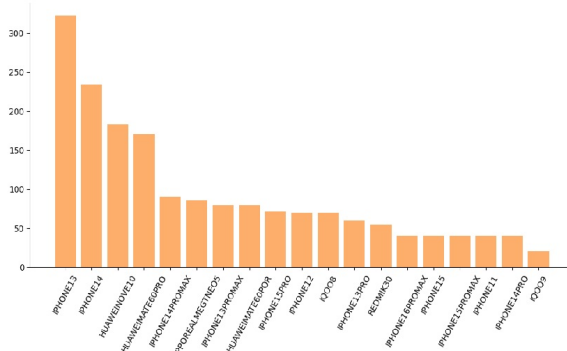


Figure 4: Statistics on the image capture equipment.

A.2 Metrics

Here, we provide more details about the metrics used in the main manuscript.

The **Accuracy** metric quantifies the proportion of questions where the predicted answer precisely corresponds with any of the designated target answers for that question.

For the Average Normalized Levenshtein Similarity (**ANLS**), we follow previous works, which is defined as follows:

$$\text{ANLS} = \frac{1}{N} \sum_{i=0}^N (\max_j s(a_{ij}, o_{qi})), \quad (1)$$

where $s(a_{ij}, o_{qi})$ is defined as $1 - NL(a_{ij}, o_{qi})$ when $NL(a_{ij}, o_{qi})$, the normalized Levenshtein distance, is less than a predefined threshold τ ; otherwise 0. we set the threshold $\tau = 0.5$, as previous works do.

In the main manuscript, we introduce the **Consistency score**, a robustness metric designed to assess the resilience of models when handling the same question across images captured under various conditions. This metric calculates the document-level accuracy; a model’s response is considered accurate only if it correctly answers the question in all four distinct scenarios presented.

A.3 Case Study

To clearly illustrate the gap between real-world captured and scanned/digital document images, and to thoroughly analyze the performance differences in these two scenarios, here, we provide several

examples from the origin benchmark and WildDoc, along with the answers of the leading MLLM, Qwen2.5-VL.

As shown in Figure 5, Qwen2.5-VL-72B correctly answers the question in the original DocVQA [30] benchmark, because the scanned/digital document images are clear and well-aligned. In contrast, the model fails to answer the question correctly in WildDoc, as the model incorrectly aligns cells from different rows together, and fails to locate the answer in the second example.

In Figure 6, we present two examples from the ChartQA [28] benchmark. As with the previous examples, Qwen2.5-VL encounters difficulties with real-world captured document images, which can be attributed to variations in photo angles and the presence of creases on the documents.

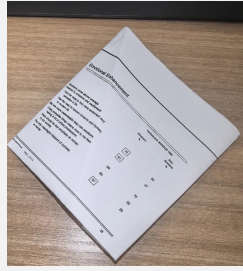
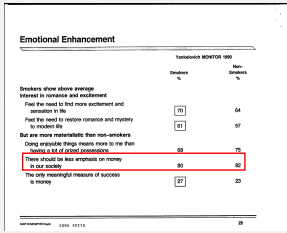
In conclusion, these cases vividly demonstrate the challenges that existing models face when dealing with real-world document images, particularly when confronted with issues such as variations in photo angles and the presence of creases in documents—issues that are seldom encountered in traditional scanned or digital document images. Consequently, conventional benchmarks often fail to reflect a model’s performance in real-world applications accurately. Our newly proposed benchmark addresses the gap, which enables a more comprehensive evaluation of a model’s ability to process complex and irregular document images.

A.4 More Information about WildDoc.

In Figure 7, we provide more examples of WildDoc. The WildDoc will be open-sourced under the CC BY-NC 4.0 License. The benchmark construction cost is mainly divided into two parts: document image acquisition and filtering. Document image acquisition costs approximately two months and \$ 5,000. The filtering session costs approximately two weeks and \$ 500. Each participant in the image capture session is provided with a detailed version of the data collection section and several data examples that we captured. For the quality inspector in the second round, it is required that they hold at least a university-level degree or higher academic qualifications, ensuring a deep level of understanding in analyzing the content of the documents (e.g., tables, infographics).

Question: What percentage of non-smokers feel there should be less emphasis on money in our society?

GT : 82

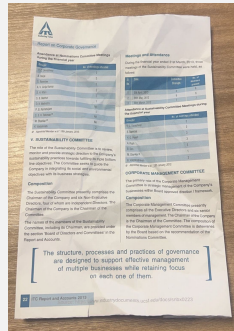
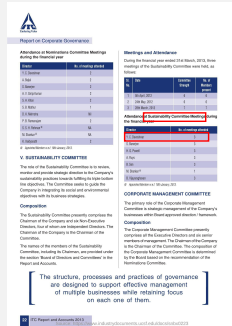


Origin: 82

WildDoc: 75

Question: How many sustainability committee meetings has Y. C. Deveshwar attended?

GT : 3



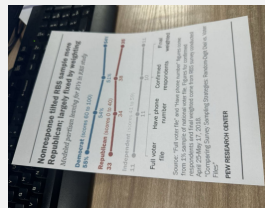
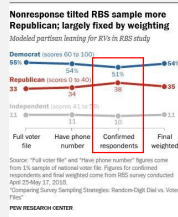
Origin: 3

WildDoc: 6

Figure 5: Evaluation results of Qwen2.5-VL-72B in the Original DocVQA [30] and our WildDoc benchmark. The answer in the figure is highlighted in red. Zoom in for the best view.

Question: What political party has the highest rate of 'confirmed respondents'?

GT : Democrat

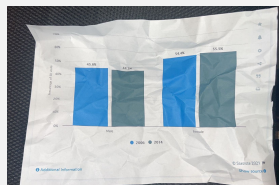
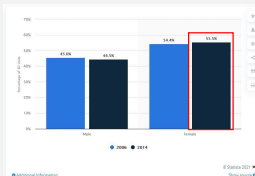


Origin: Democrat

WildDoc: Republican

Question: What percentage of emergency department visits were from female patients in 2014?

GT : 55.5%



Origin: 55.5%

WildDoc: 45.6%

Figure 6: Evaluation results of Qwen2.5-VL-72B in the Original ChartQA [28] and our WildDoc benchmark. The answer in the figure is highlighted in red. Zoom in for the best view.



Figure 7: Visualization of several examples from WildDoc.