Q-Doc: Benchmarking Document Image Quality Assessment Capabilities in Multi-modal Large Language Models

Jiaxi Huang[1], Dongxu Wu[1], Hanwei Zhu[2(✉)], Lingyu Zhu[3], Jun Xing[4], Xu Wang[5], and Baoliang Chen[1(✉)]

[1] South China Normal University, Guangzhou, China
blchen6-c@my.cityu.edu.hk
[2] Nanyang Technological University, Singapore
hanwei.zhu@ntu.edu.sg
[3] City University of Hong Kong, Hong Kong SAR, China
[4] Shenzhen Academy of Inspection and Quarantine, Shenzhen, China
[5] Shenzhen University, Shenzhen, China

Abstract. The rapid advancement of Multi-modal Large Language Models (MLLMs) has expanded their capabilities beyond high-level vision tasks. Nevertheless, their potential for **Document Image Quality Assessment (DIQA)** remains underexplored. To bridge this gap, we propose **Q-Doc**, a three-tiered evaluation framework for systematically probing DIQA capabilities of MLLMs at coarse, middle, and fine granularity levels. ***a)*** At the ***coarse level***, we instruct MLLMs to assign quality scores to document images and analyze their correlation with Quality Annotations. ***b)*** At the ***middle level***, we design distortion-type identification tasks including single-choice and multi-choice tests for multi-distortion scenarios. ***c)*** At the ***fine level***, we introduce distortion-severity assessment where MLLMs classify distortion intensity against human-annotated references. Our evaluation demonstrates that while MLLMs possess nascent DIQA abilities, they exhibit critical limitations: inconsistent scoring, distortion misidentification, and severity misjudgment. Significantly, we show that **Chain-of-Thought (CoT)** prompting substantially enhances performance across all levels. Our work provides a benchmark for DIQA capabilities in MLLMs, revealing pronounced deficiencies in their quality perception and promising pathways for enhancement. The benchmark and code are publicly available at: https://github.com/cydxf/Q-Doc

Keywords: Document Image Quality Assessment · Multi-modal Large Language Models · Multi-Level Evaluation · Chain-of-Thought Prompting

(✉) Corresponding author

1 Introduction

The rapid progress of multimodal large language models (MLLMs), such as GPT-4o [12] and open-source alternatives like DeepSeek-VL2 [24], GLM-4V [10], mPLUG-Owl3 [25], Llama-3.2-Vision [9,19] and Co-Instruct [23] has led to a significant leap in general-purpose Artificial Intelligence systems capable of jointly understanding and reasoning over visual and textual inputs. These models have demonstrated remarkable capabilities across various vision-language tasks, including visual question answering [2], image captioning [3] and visual grounding [15]. However, existing benchmarks primarily emphasize high-level semantic understanding [13,17] or general visual recognition [16,6], leaving the capacity of MLLMs in **low-level visual quality perception** [11,5], especially in task-specific domains such as document images, largely underexplored. In practice, **document image quality assessment (DIQA)** plays a pivotal role in downstream tasks like OCR, information retrieval, and document understanding [1,14,21,26]. The ability to perceive quality degradations such as defocus, underexposure, or motion blur distortion is essential for robust real-world document processing, especially in mobile capture scenarios. While recent efforts like Q-Bench [22] and BLINK [7] conduct preliminary explorations of low-level perception in natural images, and DocKylin [27] have made strides in document understanding but have not targeted quality assessment, there exists a lack of systematic benchmarking tailored to document image quality, where semantic integrity is tightly coupled with layout, text readability, and distortion sensitivity.

To bridge this gap, we propose **Q-Doc** , a comprehensive benchmark for evaluating the DIQA capabilities of MLLMs. Unlike previous works that focus on aesthetic or photographic quality in natural scenes, Q-Doc targets MLLMs' reasoning and judgment abilities on task-oriented, text-centric images captured in real-world mobile settings. Built upon the *SmartDoc-QA* dataset [20], Q-Doc includes 4,260 document images with both single and compound distortions across 30 unique documents from three practical domains: modern forms, historical letters, and receipts. As illustrated in Figure 1, our benchmark introduces a **Coarse-Middle-Fine** three-level evaluation framework that progressively dissects the quality perception ability of MLLMs across abstraction levels.

- ***Coarse-Level Quality Judgment.*** Can the MLLM provide a consistent overall quality rating (e.g., *excellent*, *poor*) for a document image, aligned with downstream OCR performance? We formulate this as a quality classification task and measure correlation with Quality Annotations using SRCC and PLCC.
- ***Middle-Level Distortion Recognition.*** Can the MLLM accurately identify the *type* of distortion (e.g., blur, lighting, focus loss) in both single-distortion (single-choice) and multi-distortion (multi-choice) settings?
- ***Fine-Level Distortion Severity Assessment.*** Can the MLLM assess the *severity* of individual distortions (e.g., mild vs. severe blur) using predefined qualitative levels (*good*/*middle*/*poor*)?

To further enhance MLLMs' reasoning process, we integrate **Chain-of-Thought (CoT)** prompting into our evaluation pipeline. By encouraging the model to explicitly reflect on text readability and its correlation with specific visual artifacts, CoT significantly improves interpretability and prediction performance across all three levels. In summary, our contributions are threefold:

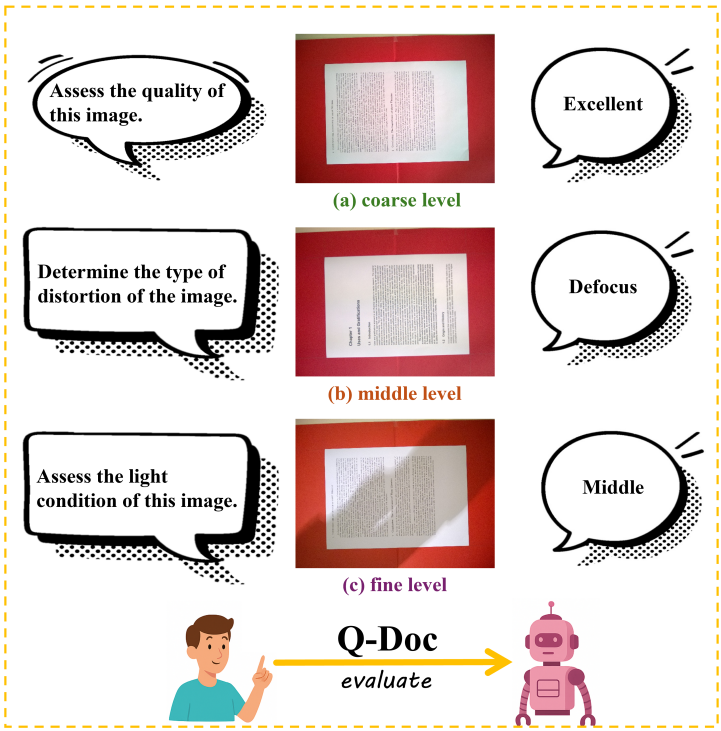

Fig. 1. Overview of the proposed Q-Doc benchmark. Q-Doc evaluates MLLMs across three levels: coarse-level quality scoring, middle-level distortion classification, and fine-level severity estimation, using real-world distorted document images.

- We introduce **Q-Doc**, a document-centric benchmark to systematically evaluate MLLMs on low-level visual perception tasks, focusing on the nuanced interplay between document distortions and textual integrity.
- We design a **three-level evaluation framework (Coarse-Middle-Fine)** that decomposes the DIQA task into interpretable subtasks, leveraging a real-world mobile document dataset with rich distortion diversity and fine-grained annotations. This framework enables systematic probing of MLLMs' perception capabilities across quality scoring, distortion classification, and severity estimation.

- We demonstrate that **CoT prompting** can meaningfully enhance MLLM performance in document image quality assessment, especially for complex multi-distortion cases.

Q-Doc establishes a new testing ground for analyzing and advancing MLLMs' real-world document understanding from a quality-centric perspective, providing insights for both foundational model development and applied document Artificial Intelligence systems.

2 Constructing the Q-Doc

2.1 General Principles

Focusing on Document-specific Visual Quality of MLLMs. Unlike general-purpose MLLM benchmarks [13,18], which evaluate broad multimodal reasoning capabilities, our proposed **Q-Doc** emphasizes a focused and layered evaluation of how MLLMs perceive and assess visual distortions in document images. Our design adheres to two core principles: (a) emphasizing the perceptual quality and textual readability of real-world document images; (b) structuring evaluation in a hierarchical fashion to progressively dissect MLLM capabilities across ***perception***, ***classification***, and ***estimation of distortion severity***.

Layered Quality Evaluation Framework. We introduce a **Coarse-Middle-Fine** three-level evaluation strategy to systematically examine different aspects of document image quality perception. This framework is tailored to simulate real-world use cases such as mobile document scanning and OCR applications. Additionally, we incorporate **CoT** prompting to investigate whether step-by-step textual reasoning improves model performance in fine-grained quality understanding.

2.2 Dataset Organization

Our dataset is curated from the *SmartDoc-QA*, a mobile-captured document image set with diverse degradation types. It consists of **4,260 real-world document images** covering 30 documents in three categories: modern documents, historical administrative letters, and receipts. Among them, 660 are single-distortion images (180 underexposed, 120 motion-blurred, 240 defocused, 120 distortion-free), while the remaining 3,600 combine multiple degradations (e.g., angle, blur, brightness).The precise distribution of distortion types is detailed in Table 1

2.3 Coarse-Level Assessment

This tier of **Q-Doc** examines whether MLLMs can accurately evaluate the **overall visual quality** of a document image.

Table 1. Distribution of Distortion Types in the Dataset.

Distortion Type	Single-Distortion	Multi-Distortion
Insufficient Brightness	180	600
Motion Blur	120	300
Defocus	240	600
Insufficient Brightness + Motion Blur	-	600
Insufficient Brightness + Defocus	-	1,200
No Distortion	120	300
Total Images	**660**	**3,600**

Task Setup. For each image, we prompt the MLLM with:

"Assess the quality of this document image and respond with ONLY one of the following words: excellent (highest quality), good (above average quality), fair (average quality), poor (below average quality), bad (lowest quality). Only respond with the single most appropriate word from the list above, with no additional text or explanation."

Each of these choices is mapped to a 5–1 scoring scale. The MLLM is expected to select a single term, which is then converted to a numerical quality prediction.

Quantitative Evaluation. We compare the model-predicted scores with the **Quality Annotations** of the same images from the SmartDoc-QA dataset. **Spearman's rank correlation coefficient (SRCC)** and **Pearson's linear correlation coefficient (PLCC)** are calculated to assess the alignment.

2.4 Middle-Level Distortion Identification

In the second tier, we assess whether MLLMs can identify the **type(s) of distortion** present in a document image.

Single Distortion (Classification Task) For each single-distortion image, we pose a 4-option multiple-choice question:

"Please determine the type of distortion in this document image: A. Insufficient brightness B. Motion blur C. Defocus D. No Distortion. Please answer the option letters directly."

We use this to evaluate the model's **distortion-type recognition accuracy** on 660 single-distortion images.

Balanced vs. Unbalanced Accuracy. Since the number of samples per distortion type in the dataset is not uniform, we report two accuracy metrics:

- **Unbalanced Accuracy** (Acc_{raw}): the standard overall accuracy computed directly over the full dataset, without adjusting for class distribution.

- **Balanced Accuracy** (Acc_{bal}): an adjusted metric to mitigate class imbalance by assigning higher weight to under-represented distortion types. Specifically, we adopt an inverse-frequency weighting scheme: for each distortion class i, let n_i be the number of test samples and acc_i the per-class accuracy. The final score is computed as:

$$Acc_{bal} = \sum_{i=1}^{C} w_i \cdot acc_i, \quad \text{where} \quad w_i = \frac{1/n_i}{\sum_{j=1}^{C} 1/n_j}, \tag{1}$$

  where C denotes the total number of distortion categories. This normalized weighting ensures that classes with fewer samples contribute more to the final accuracy, promoting fairer evaluation under class imbalance.

Multiple Distortions (Multi-label Classification Task) For each multi-distortion image, we prompt:

"Please determine the type of distortion of this document image (there may be more than one choice): A. Insufficient brightness B. Motion blur C. Defocus D. No Distortion. Please answer the option letters directly. If it's a multiple-choice question, separate the options with a comma (e.g. A, B)."

We compute multi-label classification accuracy under a **strict matching criterion**: A prediction is considered correct only if all and only the true distortion types are selected (i.e., no extra, missing, or incorrect labels). Any deviation, including over-selection, under-selection, or incorrect selection, is counted as an error.

Option Shuffling for Robustness. To prevent any positional bias in model responses, the order of distortion types is randomized across all test samples. For example, *Defocus* may appear as option C in one question but as option A in another. This ensures that the model's performance reflects true recognition capability rather than memorization of fixed option sequences.

Balanced Accuracy for Multi-label Case. As with the single-label task, we additionally report both unbalanced and balanced accuracy metrics for multi-distortion classification. Here, class-wise accuracy is computed based on exact-match correctness for each unique distortion combination. Balanced accuracy is calculated using the same inverse-frequency weighting as defined above.

2.5 Fine-Level Distortion Severity Assessment

At the fine level, we evaluate whether MLLMs can judge the **severity of each distortion type** in a document image.

Task Setup. For each distortion type, the images are manually annotated with one of three severity levels: *good, middle, poor*. We prompt the MLLMs with:

"Please evaluate the degree of [specific distortion] in this document image. Choose one of the following options: A. good B. middle C. poor. Answer with the letter only (A, B, or C)."

This task is run for each distortion type independently. The model's predictions are compared against human-annotated labels to compute accuracy.

To ensure precise evaluation, we adopt a stratified query strategy. For **single-distortion images**, we group the samples by distortion type and ask MLLMs to estimate the severity of that particular distortion. For example, given an image annotated with underexposure, we prompt the model to assess the degree of *brightness degradation* only.

For **multi-distortion images**, each sample contains both brightness-related and blur-related degradation. To disentangle these effects, we prompt the model twice per image: once to assess the severity of brightness degradation, and once to assess the severity of blur. These responses are evaluated separately for each distortion dimension.

It is noteworthy that while the blur category in multi-distortion images may encompass both defocus and motion-related degradations, we intentionally focus the severity assessment on defocus blur. Compared to motion blur, defocus exhibits more stable and perceptually distinct severity gradations, making it a more reliable target for fine-grained quality evaluation. This design choice enables consistent supervision and facilitates clearer performance comparison across models, while avoiding ambiguity introduced by temporally dynamic artifacts.

2.6 Chain-of-Thought Prompting for Performance Boost

To improve interpretability and quality estimation, we introduce **CoT** prompting strategies, as recent work has shown that CoT significantly enhances performance in MLLMs [8,4,28]. For instance, in the coarse-level task, we modify the prompt to:

"Assess the quality of this document image and respond with ONLY one of the following words: excellent (highest quality), good (above average quality), fair (average quality), poor (below average quality), bad (lowest quality). ***Analyze the impact it may have on OCR recognition, thinking about the following questions: Is the image clear (is the edge of the text sharp)? Is the light uniform? Do shadows or reflections obscure the text? Based on the above analysis, give the final image quality evaluation results.*** *Only respond with the single most appropriate word from the list above, with no additional text or explanation."* Figure 2 visually highlights how textual clarity and lighting awareness are integrated into step-by-step reasoning.

Ablation Study. We compare model performance with and without CoT across all three levels. Results show *significant accuracy improvement*, especially at the middle level multi-distortion identification, suggesting that guided reasoning about text readability helps MLLMs better understand visual degradation in documents.

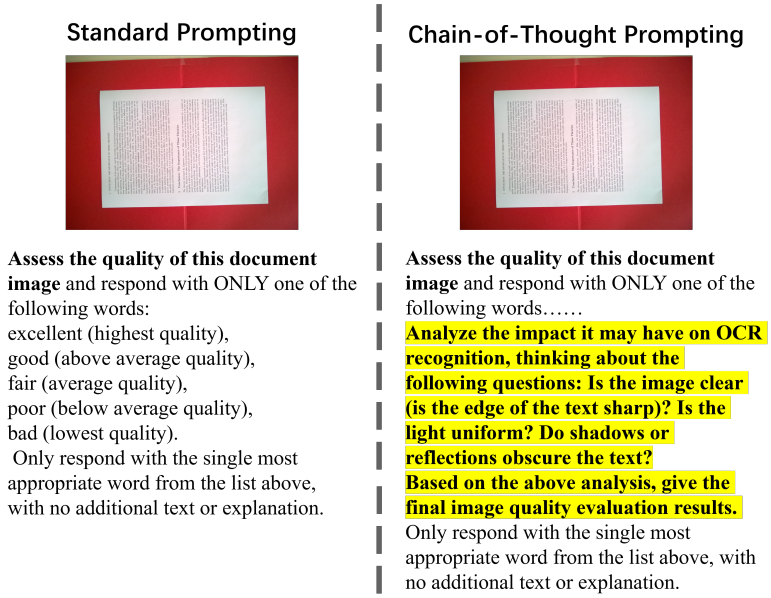

Fig. 2. Illustration of Chain-of-Thought (CoT) prompting. The model is encouraged to reason about textual readability and document clarity before producing a quality judgment.

3 Results on Q-Doc

We evaluate six representative MLLMs on our proposed three-level Q-Doc. Each model is tested under zero-shot settings, including both competitive open-source MLLMs (Co-Instruct, Llama-3.2-11B-Vision-Instruct [19], mPLUG-Owl3-7B-241101, GLM-4V-9B, DeepSeek-VL2) and one leading commercial model (GPT-4o-241120). The experiment follows our Coarse-Middle-Fine framework and is aligned with OCR-grounded quality signals. All tasks are evaluated with both unbalanced and balanced accuracy metrics when applicable. As shown in Fig. 3, we visualize the overall performance of all MLLMs across six key quality metrics in our benchmark. This provides a quick comparison of model capabilities at each level of evaluation.

3.1 Coarse-Level: Quality Scoring Correlation with Quality Annotations

In the Coarse-Level task, we measure how well MLLMs can assess the *overall visual quality* of document images. Each model's output is mapped to a 5-point scale, and its correlation with Quality Annotations is reported using Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC).

Table 2. Coarse-Level: SRCC and PLCC between MLLM-predicted quality and Quality Annotations. The best results are highlighted in bold, and the second-best results are underlined.

Model	SRCC	PLCC
Co-Instruct	0.3840	0.2686
Llama-3.2-11B-Vision-Instruct	0.2122	0.2069
mPLUG-Owl3-7B-241101	0.3607	0.2879
GLM-4V-9B	0.2162	0.1821
DeepSeek-VL2	<u>0.4474</u>	<u>0.3713</u>
GPT-4o-241120 *(closed-source)*	0.1321	0.0610
DeepSeek-VL2 + CoT	**0.4603**	**0.3886**

Observations. Table 2 shows that open-source models such as DeepSeek-VL2 and mPLUG-Owl3 show strong correlation with Quality Annotations, demonstrating their quality-awareness in document perception. Surprisingly, GPT-4o performs significantly worse, despite being a state-of-the-art commercial model. Notably, DeepSeek-VL2 combined with Chain-of-Thought (CoT) prompting yields a substantial performance gain, highlighting the value of reasoning-focused guidance in scoring document quality.

3.2 Middle-Level: Distortion Type Recognition

This level evaluates a model's ability to identify present distortions, either single (via single-choice) or multiple (via multi-choice), in a document image. To account for label imbalance, we report both unbalanced and balanced accuracy scores.

Table 3. Middle-Level: Accuracy (%) on distortion classification. The best results are highlighted in bold, and the second-best results are underlined.

Model	Single Distortion		Multiple Distortion	
	Acc_{raw}	Acc_{bal}	Acc_{raw}	Acc_{bal}
Co-Instruct	25.08	12.13	5.35	1.02
Llama-3.2-11B-Vision-Instruct	33.74	<u>55.80</u>	15.57	7.78
mPLUG-Owl3-7B-241101	18.54	2.95	6.00	4.20
GLM-4V-9B	28.42	30.69	15.76	9.84
DeepSeek-VL2	<u>41.19</u>	31.23	24.61	33.68
GPT-4o-241120 *(closed-source)*	38.45	**56.50**	<u>34.54</u>	**62.54**
DeepSeek-VL2 + CoT	**45.74**	33.26	**36.24**	<u>51.20</u>

Observations. As detailed in Table 3, most MLLMs perform significantly better on single-distortion detection than on the more challenging multi-distortion

classification task. GPT-4o achieves the highest balanced accuracy in both categories, while DeepSeek-VL2 combined with CoT demonstrates the most consistent overall performance.

It is worth noting that the accuracy gaps in the multi-distortion task appear larger than in other subtasks. This is primarily due to our **strict evaluation criterion**: a model's prediction is considered correct only if it fully matches the ground truth set of distortion types. Any over-selection, under-selection, or incorrect selection leads to a wrong prediction. This design emphasizes the model's ability to perform precise multi-label classification, but it also makes the task particularly unforgiving, especially when multiple distortions co-occur. Therefore, while the absolute accuracy values may seem low, they still reflect meaningful distinctions in MLLM capability under a high-precision standard.

3.3 Fine-Level: Severity Estimation

At this level, we evaluate whether the MLLMs can correctly estimate the severity of visual distortions *(good, middle, poor)*. The task is conducted on both single- and multi-distortion images.

Table 4. Fine-Level: Accuracy (%) on distortion severity estimation. The best results are highlighted in bold, and the second-best results are underlined.

Model	**Single Distortion**		**Multi Distortion**
	Acc_{raw}	Acc_{bal}	Overall Acc.
Co-Instruct	22.62	17.23	48.08
Llama-3.2-11B-Vision-Instruct	25.19	38.73	36.99
mPLUG-Owl3-7B-241101	<u>55.33</u>	34.65	51.63
GLM-4V-9B	43.68	<u>45.65</u>	33.49
DeepSeek-VL2	35.14	33.68	<u>54.71</u>
GPT-4o-241120 *(closed-source)*	**68.89**	**50.86**	47.99
DeepSeek-VL2 + CoT	34.69	35.31	**57.58**

Observations. According to Table 4, GPT-4o achieves the highest accuracy on single-distortion severity estimation, while DeepSeek-VL2 shows the best performance on multi-distortion images. Interestingly, several models, including DeepSeek-VL2 and its CoT-enhanced variant, perform even better on multi-distortion samples than on single-distortion ones.

This result may appear counter-intuitive. However, it stems from differences in task framing rather than evaluation looseness. While both settings evaluate the same two distortion dimensions (brightness and blur) the distribution of distortion subtypes differs across single- and multi-distortion groups. Specifically, our multi-distortion evaluation focuses on well-calibrated and perceptually distinguishable subtypes (e.g., gradual underexposure and controlled defocus), which facilitate more stable assessment conditions. In contrast, single-distortion

images include a broader spectrum of sub-variations, potentially increasing intra-class ambiguity.

This design choice reflects our intent to balance task challenge with annotation consistency. By framing multi-distortion severity estimation around more uniformly gradable distortions, we ensure that performance gains reflect model capability rather than dataset variance, thereby reinforcing the diagnostic value of fine-level assessment.

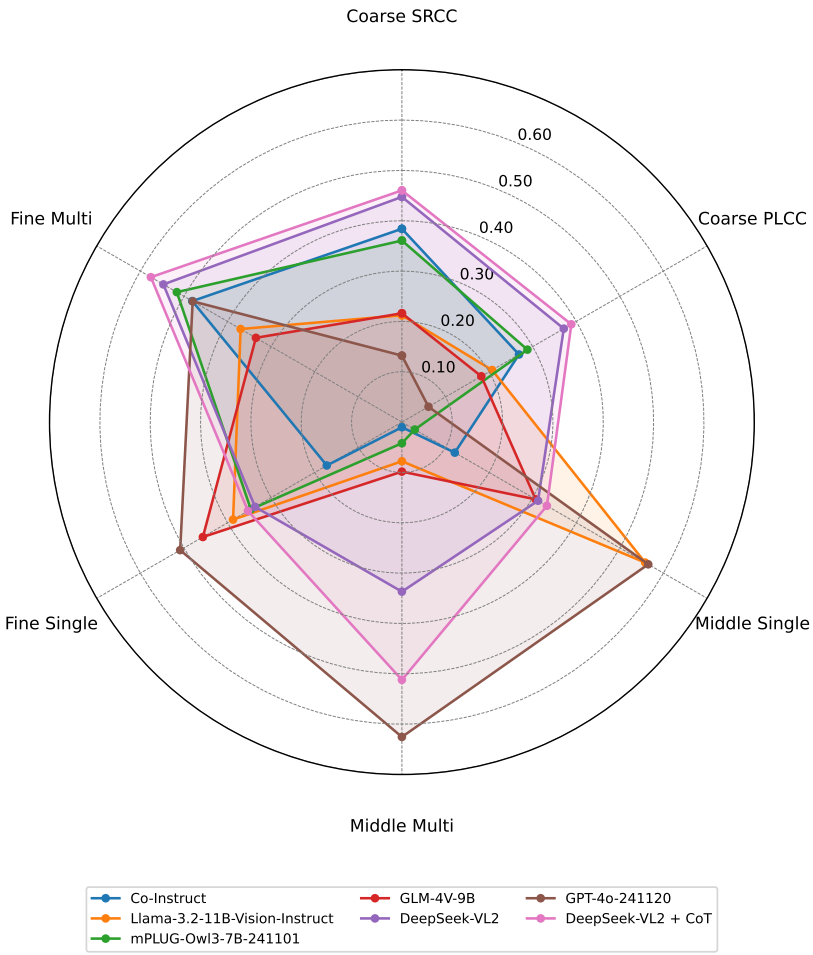

Fig. 3. Radar chart of MLLM performance across six quality evaluation metrics: Coarse-level correlation (SRCC, PLCC), Middle-level balanced accuracy (Single/Multiple), and Fine-level balanced accuracy (Single/Multiple).

3.4 Insights on MLLM Performance

While our evaluation involves six representative MLLMs, our core objective is to understand how these general-purpose models perform across the three hierarchical layers of document image quality perception.

Our findings reveal several important observations:

- **Performance inconsistency across levels.** For most models, quality perception ability is *not consistently strong across all levels*. For example, Llama-3.2-11B-Vision-Instruct shows excellent performance in middle-level distortion classification, especially on single distortions, but performs poorly in both coarse-level quality scoring and fine-level severity estimation. In contrast, mPLUG-Owl3-7B-241101 shows decent performance in fine- and coarse-level tasks estimation, yet falls behind significantly in middle-level distortion identification. Even GPT-4o-241120, which excels at both middle-level classification and fine-level granularity, surprisingly underperforms in overall quality scoring, indicating a potential mismatch between high-resolution perception and global document quality judgment.
- **DeepSeek-VL2 exhibits the most balanced capability.** Among all models evaluated, DeepSeek-VL2 is the only one that consistently maintains competitive performance across Coarse, Middle, and Fine levels. This balance suggests a more unified and generalizable perception capability, potentially attributable to its **Mixture-of-Experts (MoE)** architecture. Such architecture may allow the model to dynamically activate specialized submodules for different granularities of visual reasoning, enabling both holistic and localized understanding.
- **Level-specific strengths hint at architectural biases.** The inconsistency of most models across evaluation layers likely stems from architectural or training-phase biases. For example, models like GPT-4o-241120 appear more sensitive to local patterns (reflected in decent fine-level performance) but lack global context alignment (shown in coarse-level correlation drops). These variations suggest that visual token integration, prompt understanding, and training corpus balance may all shape a model's quality perception profile.
- **Chain-of-Thought prompting enhances multi-step visual reasoning.** Across all levels, CoT-enhanced DeepSeek-VL2 outperforms its vanilla counterpart, particularly in the Middle and Fine levels. This demonstrates that even for tasks requiring visual-textual alignment, explicit reasoning guidance significantly improves reliability and interpretability of MLLM predictions.

In summary, our results indicate that current MLLMs exhibit *fragmented* strengths in document quality assessment, where some models excel in isolated subtasks but lack cross-layer consistency. DeepSeek-VL2, supported by its modular design, stands out as a promising direction toward building quality-aware MLLMs with generalizable and layered perception capabilities.

4 Conclusion

In this study, we propose **Q-Doc**, a comprehensive three-level benchmark designed to evaluate the DIQA abilities of MLLMs. Unlike existing general-purpose benchmarks, **Q-Doc** explicitly decomposes DIQA into interpretable subcomponents, quality scoring, distortion recognition, and severity estimation, each grounded in real-world document imaging challenges. Our extensive experiments confirm that while today's MLLMs show promise in OCR-aware perception, they still struggle with distortion interpretation and quantification. We also validate that CoT prompting can significantly improve their diagnostic capabilities. We hope **Q-Doc** inspires future development of quality-aware, robust, and readable document understanding systems based on MLLMs.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant 62401214), in part by the National Natural Science Foundation of China (Grant 62371310), in part by Shenzhen Science and Technology Program (JCYJ20241202124415021).

References

1. Alaei, A., Bui, V., Doermann, D., Pal, U.: Document image quality assessment: a survey. ACM computing surveys **56**(2), 1–36 (2023)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
4. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1254–1262 (2024)
5. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3677–3686 (2020)
6. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), `https://arxiv.org/abs/2306.13394`
7. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
8. Ge, J., Luo, H., Qian, S., Gan, Y., Fu, J., Zhang, S.: Chain of thought prompt tuning in vision language models. arXiv preprint arXiv:2304.07919 (2023)
9. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
10. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500 (2024)

11. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
13. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
14. Li, H., Zhu, F., Qiu, J.: Cg-diqa: No-reference document image quality assessment based on character gradient. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3622–3626. IEEE (2018)
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
16. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
17. Liu, Z., Fang, F., Feng, X., Du, X., Zhang, C., Wang, N., Zhao, Q., Fan, L., GAN, C., Lin, H., et al.: Ii-bench: An image implication understanding benchmark for multimodal large language models. Advances in Neural Information Processing Systems **37**, 46378–46480 (2024)
18. Lu, J., Rao, J., Chen, K., Guo, X., Zhang, Y., Sun, B., Yang, C., Yang, J.: Evaluation and mitigation of agnosia in multimodal large language models. arXiv preprint arXiv:2309.04041 **3** (2023)
19. Meta: Llama-3.2-11b-vision-instruct. `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct` (2024)
20. Nayef, N., Luqman, M.M., Prum, S., Eskenazi, S., Chazalon, J., Ogier, J.M.: Smartdoc-qa: A dataset for quality assessment of smartphone captured document images-single and multiple distortions. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1231–1235. IEEE (2015)
21. Van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks (2020)
22. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., Lin, W.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. In: ICLR (2024)
23. Wu, H., Zhu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Li, C., Wang, A., Sun, W., Yan, Q., et al.: Towards open-ended visual quality comparison. In: European Conference on Computer Vision. pp. 360–377. Springer (2024)
24. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
25. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840 (2024)
26. Ye, P., Doermann, D.: Document image quality assessment: A brief survey. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 723–727. IEEE (2013)

27. Zhang, J., Yang, W., Lai, S., Xie, Z., Jin, L.: Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9923–9932 (2025)
28. Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., Yang, Y.: Improve vision language model chain-of-thought reasoning. arXiv preprint arXiv:2410.16198 (2024)