Evaluating Vision-Language Models for Scanned Image Quality Assessment

□ Ali Haider, Sung-Ho Bae / Kyung Hee University

Abstract

Image quality assessment (IQA) supports restoration, perception, and document analysis, but remains challenging under scanning distortions. In this work, we evaluate open-source vision language models Qwen-VL, Gemma-3 Vision, and LLaVA for scanned image quality assessment. Using the Descan-18k dataset with synthetic artifacts such as bleed through, scanner streaks, halftoning, and blur, we prompt each model to grade images as Good, Fair, or Bad. Despite standardized prompts and parsing, performance is weak: accuracy ≈ 0.33, macro-F1 ≈ 0.17, and macro precision ≈ 0.15-0.18 is low, with a strong bias toward Good and near-zero recall for Fair and Bad. These findings indicate that current models are not yet reliable graders for scanned image quality; our evaluation protocol and analyses provide a reproducible baseline and point to calibration or task specific training for practical use.

I . Introduction

Image Quality Assessment (IQA) [1] is a fundamental problem in computer vision that aims to quantify the perceptual and structural degradation of an image. Reliable IQA plays a critical role in downstream applications such as image restoration, enhancement, perception, and document analysis. Traditional IQA metrics such as Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) [2], and Feature Similarity (FSIM) [3] rely on low level pixel or feature correlations. While effective for controlled distortions, these approaches often fail to capture complex or content dependent degradations that influence human visual perception. Deep learning based IQA models, including DeepIQA [4], RankIQA [5], and PieAPP [6], improve correlation with human judgments by learning from annotated datasets; however, their performance is typically limited to the types of distortions present in their

training data, reducing generalization to unseen or composite degradations.

Recent advances in vision language models (VLMs) [7], [8] have introduced new opportunities for adaptive and context aware image evaluation. By jointly processing visual and textual inputs, VLMs such as CLIP [9], BLIP2 [10], Qwen-VL [7], DeepSeekVL [8], and LLaVA [11] can reason about semantic content and perceptual attributes simultaneously. This multimodal capability allows them to assess image quality via natural language instructions, offering flexibility beyond the fixed scoring schemes of traditional IQA models. Early studies report potential in aesthetic quality prediction [12], caption consistency evaluation [13], and preference alignment, suggesting that VLMs could serve as general purpose visual evaluators without task specific retraining.

A particularly challenging domain for IQA is the Descanning [14] problem, which focuses on restoring images that have undergone printing and scanning cycles. Unlike typical digital degradations, scanning introduces a complex mixture of physical and optical artifacts, including bleed through, scanner streaks, halftoning, desaturation, color misalignment, geometric warping, and surface texture flattening [14]. These distortions are often spatially correlated and context dependent, making them difficult to quantify using traditional pixel based metrics. Figure 1 illustrates representative examples across quality levels, showing how noise, misalignment, and streaking progressively alter visual structure and perceptual quality.

In this work, we investigate whether modern open source VLMs can serve as reliable graders for scanned image quality. We evaluate Qwen-VL [7], Gemma-3 [15], and LLaVA [11] on the Descan-18k dataset, which contains images degraded by synthetically induced scan like artifacts at multiple severity levels. Each model is prompted to assign an ordinal quality label Good, Fair, or Bad based on perceptual clarity and structural integrity. Our experiments reveal limited performance, with low accuracy and F1 scores and a strong bias toward the Good class, indicating that current VLMs are not yet reliable graders for scanned

<Figure 1> Examples of synthetic scanning distortions across quality levels. The Good image remains largely clean and readable. The Fair and Bad images show increasing shading and wear, scanner streaks and banding, line artifacts, bleed through, and mild misalignment, which together degrade structure and legibility.

image quality without additional calibration or task specific training.

Ⅱ. Methods

This work evaluates whether modern vision language models can act as perceptual graders for scanned images affected by a range of visual distortions. The pipeline integrates dataset preparation, synthetic distortion generation, prompt design, model inference, and quantitative evaluation within a unified framework.

1. Dataset and distortion generation

We use the Descan-18k dataset, which contains clean images and images degraded to simulate real-world scanning artifacts. To mimic degradation introduced by printing and scanning, we apply synthetic distortions that include bleed through, scanner streaks and banding, halftoning, Gaussian and motion blur, geometric misalignment and skew, color desaturation and tone shift, and texture flattening with background noise. Each distortion is parameterized at three severity levels aligned with the target labels Good, Fair, and Bad. For every clean image, multiple distorted variants are generated to span the space of artifact types and intensities. Figure 1 illustrates representative examples that show how progressive degradation alters structure and perceptual quality.

2. Vision language model selection

We evaluate open source multimodal models to cover a range of capabilities. The primary models are Qwen-VL, Gemma-3, and LLaVA, used from public checkpoints without fine tuning. Each combines a visual encoder with a large language model for reasoning and text generation. Running all models in zero-shot isolates their inherent visual understanding rather than dataset specific learning.

3. Prompt design and inference strategy

We use a single canonical instruction to elicit deterministic, one-token judgments:

“You are an expert image quality evaluator. Carefully examine the image and rate its quality. Consider clarity, sharpness, alignment, color consistency, and artifacts such as bleed-through, streaks, halftoning, blur, or noise. Respond with one word only: Good, Fair, or Bad.”

Semantically equivalent variants such as “Grade the quality of this image” and “Rate this image as Good, Fair, or Bad” are also tested to check prompt sensitivity. Each image is processed independently. Model outputs are parsed and normalized to the three target labels, and off-format responses are filtered with simple keyword rules and normalization.

4. Evaluation Metrics and results

We evaluate LLaVA, Gemma-3 Vision, and Qwen-VL on Descan-18k using confusion-matrix-derived metrics. We report accuracy, macro-F1, and macro precision, and we summarize severity-binned confusion matrices to show how errors change with distortion. All runs are zero-shot with the same

<Table 1> Overall accuracy, macro-F1, and macro precision on Descan-18k for LLaVA, Gemma-3 Vision, and Qwen-VL under the standardized prompt

Model	Accuracy	Macro-F1	Macro Precision
LLaVA	0.3333	0.1667	0.1111
Gemma-3	0.3296	0.1686	0.1769
Qwen-VL	0.3278	0.1681	0.1577

instruction and decoding settings.

These results are uniformly weak and similar across models, indicating poor balance across classes and limited sensitivity to increasing distortion severity.

Ⅲ. Implementation Details

Experiments use Python with Hugging Face Transformers on PyTorch, using vLLM as the inference engine for fast, memory-efficient generation. We run zero-shot inference for Qwen-VL, Gemma-3, and LLaVA on NVIDIA RTX 4090 GPUs in mixed precision. Inputs are resized to 512×512 and normalized by each model's native processor. A single canonical instruction is used with deterministic decoding. Predictions are normalized to {Good, Fair, Bad}.

Ⅳ. Conclusion

On Descan-18k, LLaVA, Gemma-3 Vision, and Qwen-VL perform weakly as zero-shot graders of scanned image quality: accuracy is near one third and macro-F1 is low, reflecting bias toward Good and poor detection of Fair and Bad. These results indicate current open VLMs are not reliable for this task without additional calibration or training; practical next steps include class-prior reweighting, temperature scaling, few-shot exemplars, and lightweight adapters to improve class balance and robustness.

References

[1] C. Ma, Z. Shi, Z. Lu, S. Xie, F. Chao, and Y. Sui, "A Survey on Image Quality Assessment: Insights, Analysis, and Future Outlook," Feb. 12, 2025, *arXiv*: arXiv:2502.08540. doi: https://doi.org/10.48550/arXiv.2502.08540.

[2] J. Nilsson and T. Akenine-Möller, "Understanding SSIM," June 29, 2020, *arXiv*: arXiv:2006.13846. doi: https://doi.org/10.48550/arXiv.2006.13846.

[3] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011, doi: https://doi.org/10.1109/TIP.2011.2109730.

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Trans. on Image Process.*, vol. 27, no. 1, pp. 206-219, Jan. 2018, doi: https://doi.org/10.1109/TIP.2017.2760518.

[5] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from Rankings for No-Reference Image Quality Assessment," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 1040-1049. doi: https://doi.org/10.1109/ICCV.2017.118.

[6] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual Image-Error Assessment Through Pairwise Preference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, June 2018, pp. 1808-1817. doi: https://doi.org/10.1109/CVPR.2018.00194.

References

[7] J. Bai et al., "Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities," Aug. 24, 2023, *arXiv*: arXiv:2308.12966. doi: https://doi.org/10.48550/arXiv.2308.12966.

[8] H. Lu et al., "DeepSeek-VL: Towards Real-World Vision-Language Understanding," Mar. 11, 2024, *arXiv*: arXiv:2403.05525. doi: https://doi.org/10.48550/arXiv.2403.05525.

[9] H.-Y. Chen et al., "Contrastive Localized Language-Image Pre-Training," Feb. 19, 2025, *arXiv*: arXiv:2410.02746. doi: https://doi.org/10.48550/arXiv.2410.02746.

[10] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," June 15, 2023, *arXiv*: arXiv:2301.12597. doi: https://doi.org/10.48550/arXiv.2301.12597.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," Dec. 11, 2023, *arXiv*: arXiv:2304.08485. doi: https://doi.org/10.48550/arXiv.2304.08485.

[12] H. Zhou et al., "UniQA: Unified Vision-Language Pre-training for Image Quality and Aesthetic Assessment," Oct. 2024, Accessed: Oct. 08, 2025. [Online]. Available: https://openreview.net/forum?id=8mE8KNHTjd

[13] Q. Ye, X. Zeng, F. Li, C. Li, and H. Fan, "Painting with Words: Elevating Detailed Image Captioning with Benchmark and Alignment Learning," Mar. 10, 2025, *arXiv*: arXiv:2503.07906. doi: https://doi.org/10.48550/arXiv.2503.07906.

[14] J. Cha et al., "Descanning: From Scanned to the Original Images with a Color Correction Diffusion Model," Feb. 08, 2024, *arXiv*: arXiv:2402.05350. doi: https://doi.org/10.48550/arXiv.2402.05350.

[15] G. Team et al., "Gemma-3 Technical Report," Mar. 25, 2025, *arXiv*: arXiv:2503.19786. doi: https://doi.org/10.48550/arXiv.2503.19786.

Authors

Ali Haider

Ali Haider received B.S. degree in Electronics Engineering from the University of Engineering and Technology, Taxila in 2019. He later served as a Research Assistant at the Center for Intelligent Systems and Security, ITU Lahore, working on deep learning for nanostructure response estimation and inverse problems. He is currently pursuing a Ph.D. in Artificial Intelligence at Kyung Hee University Global Campus, where he is a member of the Machine Learning and Visual Computing (MLVC) Lab, under the supervision of Professor Sung Ho Bae.

Sung-Ho Bae

Sung-Ho Bae (Member, IEEE) received the B.S. degree from Kyung Hee University, South Korea, in 2011, and the M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2016, respectively. From 2016 to 2017, he was a Postdoctoral Associate with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), MA, USA. Since 2017, he has been an Assistant Professor with the Department of Computer Science and Engineering, Kyung Hee University. He has been involved in model compression/interpretation for deep neural networks and inverse problems in image processing and computer vision.