PCRI: Measuring Context Robustness in Multimodal Models for Enterprise Applications

Hitesh Laxmichand Patel, Amit Agarwal, Srikant Panda, Hansa Meghwani, Karan Dua, Paul Li, Tao Sheng, Sujith Ravi, Dan Roth

Oracle AI

Correspondence: hitesh.laxmichand.patel@oracle.com

Abstract

The reliability of Multimodal Large Language Models (MLLMs) in real-world settings is often undermined by sensitivity to irrelevant or distracting visual context, an aspect not captured by existing evaluation metrics. We introduce the **Patch Context Robustness Index (PCRI)**, the first systematic and interpretable score for quantifying MLLM robustness to variations in visual context granularity, measuring performance changes between localized image patches and full-image input.

Applying PCRI to 19 state-of-the-art MLLMs across 15 vision-language benchmarks, we find that most leading models remain brittle to background noise, with only a few, such as InternVL2-26B and Qwen2VL-72B, demonstrating consistent robustness across tasks. PCRI analysis also highlights how different model architectures handle and integrate visual context, offering actionable diagnostic insight for both researchers and practitioners.

PCRI enables rigorous comparison of context robustness, supporting principled model selection and guiding the development of future architectures and training strategies for robust, real-world deployment.

1 Introduction

MLLMs have rapidly transformed real-world applications such as visual question answering (Pattnayak et al., 2024, 2025), e-commerce product search (Meghwani et al., 2025; Singh, 2023, 2021), interactive assistants, document understanding, (Patel et al., 2024, 2025; Agarwal et al., 2025a,c), accessibility for visually impaired users (Panda et al., 2025a,b,c), and synthetic data-pipelines (Dua et al., 2025; Agarwal et al., 2024a,b; Singh, 2022). In these deployments, models must reliably extract relevant cues from complex visual scenes, for example, correctly identifying a product despite background clutter or assisting visually impaired users in noisy environments, to support safety, fairness, & user experience. However, despite impressive progress in academic benchmarks, current MLLMs often fail to generalize when exposed to complex, noisy, or dynamic visual environments.

Current evaluation protocols typically measure model performance on static, full-image contexts, implicitly assuming uniform relevance of all visual regions or relying on model's capability to filter the relevant information to solve a given task. This assumption rarely holds in practice: real-world images often contain clutter, occlusions, or irrelevant backgrounds that can mislead even advanced models. In contrast, we explicitly evaluate model behavior on both full images and localized image patches to examine the sensitivity to visual context granularity. Recent studies have documented failures due to missed local details (Zhang et al., 2024, 2023), fragmentation from cropping (Zhu et al., 2024; Ma et al., 2024), and performance drops under visual perturbations (Qiu et al., 2024). Such failures have direct implications for real-world reliability, user trust, and downstream decision making.

Practitioners and system designers need tools to quantify and compare the context robustness of MLLMs, defined as the ability of the models to maintain performance when the visual scene changes in granularity or distractor content, enabling informed model selection, and mitigation of hidden failure modes. However, to our knowledge, no standardized metric or score currently exists to quantify this form of robustness.

In this work, we introduce Patch Context Robustness Index (PCRI), a novel practical score to measure the sensitivity of MLLMs to variations in visual context granularity. PCRI directly quantifies performance differences when models process localized image patches versus full-image contexts. Our contributions are as follows:

- We propose PCRI, the first quantitative score

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 195–214

designed specifically to measure the context robustness of MLLMs under varying visual granularities.

- We present a structured, patch-based evaluation framework to systematically understand sensitivity to visual context in MLLMs.
- We present a large-scale study of 19 state-of-the-art MLLMs on 15 vision-language datasets, revealing significant and previously unmeasured context sensitivity.

Our evaluation shows that even leading MLLMs remain surprisingly brittle to context variation, with only a few architectures demonstrating robust, human-like reasoning.

2 Related Work

Robustness Benchmarks for Multimodal Models. The increasing adoption of MLLMs in real-world applications has driven extensive research on their robustness to input variations. Recent works have constructed challenging benchmarks to probe model reliability under diverse perturbations (Agarwal et al., 2025b). Qiu et al. (2024) (MMRobustness) systematically evaluate MLLMs on distribution shifts via 17 image and 16 text perturbation techniques, introducing metrics like MultiModal Impact (MMI) and Missing Object Rate (MOR). Similarly, R-Bench (Li et al., 2025) targets real-world corruptions by modeling the complete imaging pipeline, including in-the-wild and machine-induced distortions across 33 dimensions, and proposes comprehensive robustness evaluations for 20 MLLMs. Both studies highlight the vulnerability of MLLMs to common and complex perturbations, yet focus primarily on distribution shift and absolute/relative performance drops under corruptions.

Task-Specific and Contextual Robustness. Beyond generic robustness, certain tasks probe more nuanced forms of context sensitivity. For example, VCR (Zhang et al., 2025) challenges vision-language models to restore occluded embedded text, requiring pixel-level reasoning about local and global context. While such tasks advance the frontier of context-aware modeling, their evaluations are task-specific and do not yield general-purpose robustness metrics or score.

Attention Mechanisms and Visual Context in MLLMs. Numerous studies investigate the failure modes of attention mechanisms in MLLMs, revealing sensitivity to object size, distractors, and spatial arrangement (Zhang et al., 2024, 2023). Architectural innovations, such as multi-resolution encoding (Ma et al., 2024; Zhu et al., 2024; Thapa et al., 2024), token pruning (Chen et al., 2024a), and text-relevant patch selection (Ye et al., 2024) have been developed to mitigate context distractions, but typically optimize for efficiency or accuracy without providing standardized, interpretable measures of context robustness.

Summary and Our Contribution. In summary, while recent benchmarks and metrics have significantly advanced the evaluation of MLLMs under distributional shift, corruption, and task-specific complexity, there remains a critical gap: *no prior work provides a unified, score-driven framework for quantifying MLLM robustness to visual context granularity across diverse tasks and architectures*. PCRI fills this gap, offering a standardized, interpretable, and broadly applicable score for evaluating and comparing the context robustness of MLLMs, and enabling more principled model selection and deployment in practice.

3 Methodology

Our goal is to systematically evaluate whether MLLMs can reason robustly over both localized and global visual contexts, an essential capability for real-world deployment, where distractors and irrelevant background are common. Existing evaluation protocols rarely measure this aspect, and ad hoc approaches (such as object-centric cropping) introduce biases and are impractical at scale. We therefore introduce a simple, reproducible patch-based framework to quantify context robustness, suitable for diverse models and tasks.

Patch-Based Evaluation Framework. Given an image, we partition it into $n \times n$ non-overlapping, equally sized patches, with n controlling the granularity. For each patch, the model is evaluated independently, isolating its ability to extract information from other patches(see Figure 1). A regular grid ensures unbiased, interpretable, and systematic analysis, avoiding the pitfalls of object-centric or saliency-based methods, and enabling direct comparison across models and datasets. We evaluate at three granularities:

- **Full-Image Context ($n = 1$):** The model receives the entire image.
- **Moderate Granularity ($n = 2$):** The model receives each 2×2 patch independently.

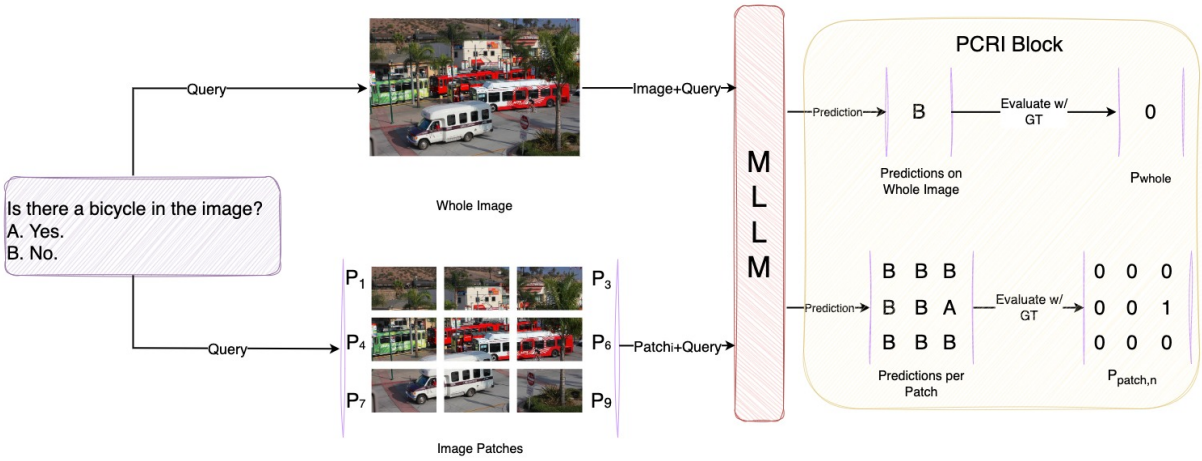

Figure 1: PCRI-based evaluation framework. An MLLM processes a query using either the full image (top) or individual patches (bottom). Predictions are compared against ground truth to compute PCRI, assessing robustness to context variations in multimodal reasoning.

- **Fine Granularity ($n = 3$):** The model receives each 3×3 patch independently.

Larger n were explored in ablations (Appendix A.1 & A.4.6), with diminishing returns and substantially increased computational cost (n^2 evaluations per image).

Patch Context Robustness Index (PCRI). PCRI quantifies a model's sensitivity to changes in visual context granularity. Formally, for a given n:

$$\text{PCRI}_n = 1 - \frac{P_{\text{patch},n}}{P_{\text{whole}}} \tag{1}$$

where:

- $P_{\text{patch},n}$ is the maximum performance achieved (per sample) over all $n \times n$ patches.
- P_{whole} is the performance with the full image.

Aggregation policy (max over patches). We aggregate patch scores with a *max* operator because PCRI diagnoses whether global context distracts the model. Patches provide minimal-context views; contrasting the best local patch with the full image reveals if access to global context helps or hurts. If the best patch rivals or exceeds full-image performance, the model is likely relying on spurious global cues (global context as a distractor); if it lags, the task or model benefits from global integration. Averaging over patches would dilute informative regions with many irrelevant ones and obscure this distraction signal (see Appendix A.1).

Interpretation. PCRI is a comparative score, agnostic to indiviudal metrics and dataset, that captures a model's sensitivity to visual context granularity. Table 1 summarizes the key scenarios.

PCRI Value	Interpretation
≈ 0	Model is robust; performs equally well on full image and patches.
< 0	Global context distracts; model harmed by irrelevant background.
$> 0, \leq 1$	Model needs global context to solve the task; patch input omits necessary information.
$\ll 0$ or undefined	$P_{\text{whole}} \to 0$; model cannot solve task even on the full image—interpret with caution.

Table 1: Summary of PCRI score interpretation. Each PCRI range indicates a distinct model behavior with respect to robustness against visual context.

Validity Domain & Chance. We interpret PCRI only when the full-image score is meaningfully above the dataset-specific chance floor. Let $C(d)$ denote the chance level for dataset d (e.g., $1/|\mathcal{Y}|$ for balanced $|\mathcal{Y}|$-way classification, or the documented random baseline for retrieval/captioning metrics). A model-dataset pair (m, d) is considered *valid* if:

$$P_{\text{whole}}(d, m) \geq C(d) + \Delta_{\min},$$
$$\Delta_{\min} = \max\{\delta,\, 2\,\text{SE}\}.$$

Here, SE is the standard error of P_{whole}, estimated via nonparametric bootstrap over evaluation examples, and δ is a small absolute margin on the native scale of the base metric. Unless otherwise stated, we set $\delta = 0.01$ for $[0, 1]$-scaled metrics (i.e., 1.0 percentage point). Model–dataset pairs (m, d) that fail this gate are labeled *near-chance/unstable*; in such cases, practitioners should

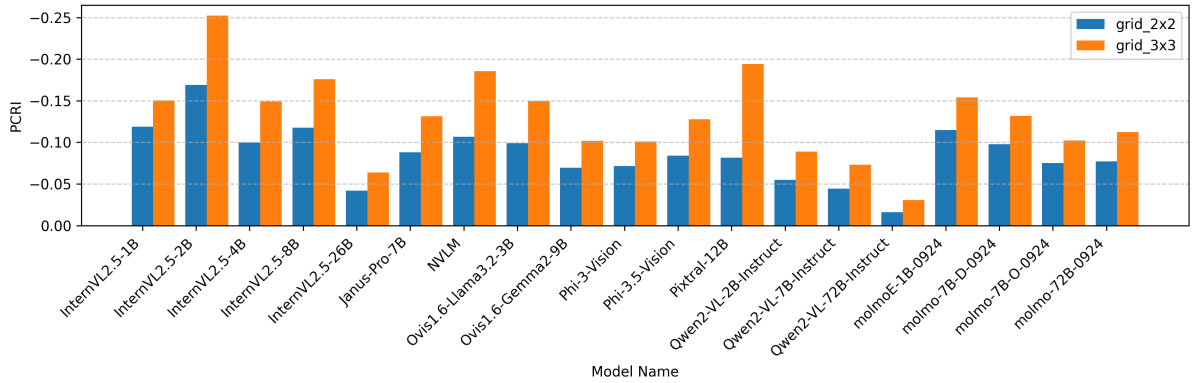

Figure 2: Avg. PCRI across 15 benchmarks for 19 MLLMs at 2×2 and 3×3 granularities. Lower PCRI values highlight model succeeds on local patches but fails on the whole image highlighting the sensitivity of the models.

not compute or interpret PCRI and may report only the underlying task metrics. In our experiments, all model–dataset pairs satisfy this gate; therefore, we report and interpret PCRI for all results. Further details can be found in Appendix A.2.

4 Experiments & Results

We evaluate PCRI on diverse models and datasets to comprehensively assess MLLM robustness to visual context granularity.

Benchmarks. We evaluate across multiple benchmarks by categorizing them into type of tasks:

- Image Captioning: MS-COCO Captions
- Multiple-Choice QA (MCQ): AI2D, BLINK, MMMU, MMStar, RealWorldQA, ScienceQA
- Yes/No Classification: AMBER, HallusionBench, MME, POPE
- Visual Question Answering (VQA): ChartQA, GQA, TextVQA, VizWiz

Models. We benchmark 19 state-of-the-art MLLMs, across various model family and sizes.

Granularity and Compute. We default to small grids $n \in \{2, 3\}$ (4–9 patches) for the best insight-per-compute; evaluation cost scales with n^2 relative to a single full-image pass. Larger n tends to fragment coherent evidence and weakens PCRI's global-context distraction probe (see Appendix A.1). For consistency & reproducibility, all evaluations use VLMEvalKit (Duan et al., 2024). Details of the datasets & models is in Appendix A.3.

4.1 Results & Discussion

We organize our analysis around core research questions central to evaluating PCRI's validity, utility & context robustness of MLLMs. Each subsection directly addresses one of these questions.

4.1.1 Do MLLMs Favor Localized Patches over Full Images?

The majority of the 19 benchmarked MLLMs exhibit negative PCRI values (Figure 2), indicating better performance on localized patches than on full images. For most models, global visual context introduces noise or distraction that outweighs its benefits for task performance. This pattern is especially pronounced in smaller InternVL variants, NVLM, and Pixtral, which show the lowest PCRI values, suggesting greater sensitivity to irrelevant context as model size or alignment decreases. Possible contributors to this trend include:

- **Visual Distraction:** Difficulty in filtering background reduces accuracy on full images.
- **Cross-Modal Attention Misalignment:** Suboptimal alignment with text prompts leads to diluted focus in global contexts.
- **Attention Overload:** Increased tokens in full images can overwhelm attention mechanisms.

In contrast, models such as InternVL2-26B & Qwen2VL-72B display near-zero PCRI, reflecting comparatively higher robustness to context variation & improved integration of global context. Overall, while a few models are closing the gap, most current MLLMs remain sensitive to context.

4.1.2 Does PCRI Align with Human-Like Contextual Reasoning?

To validate PCRI's interpretability, we conducted a human study in which three annotators solved vision-language tasks across two datasets, using both individual patches and the full image (see

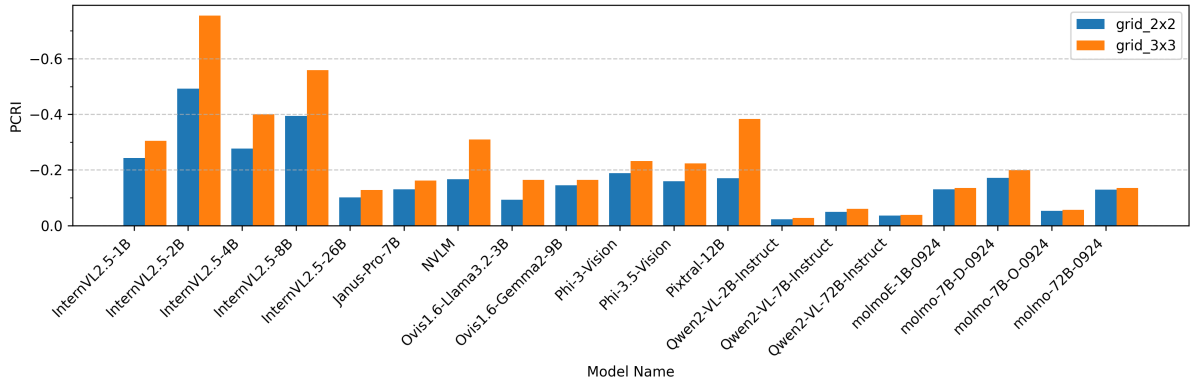

Figure 3: Avg. PCRI on MS-COCO Captioning task for different MLLMs at 2×2 and 3×3 granularities. Lower PCRI values highlight stronger performance on localized patches versus full-image contexts, notably in smaller InternVL2.5 models (<26B), NVLM and Pixtral models.

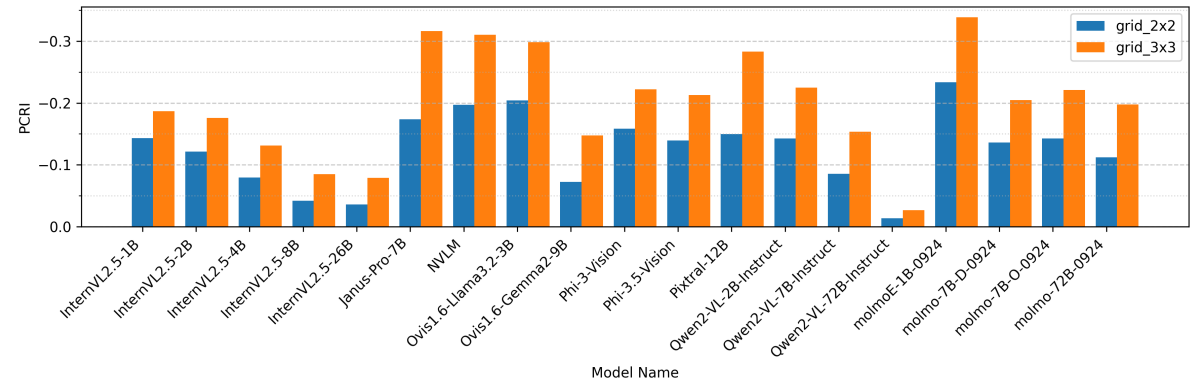

Figure 4: Avg. PCRI scores for MCQ tasks. Models exhibit moderate sensitivity to visual context granularity, with larger models (e.g., Qwen2-VL-72B) demonstrating enhanced robustness to global contextual noise, with consistent improved performance with finer patch granularities (3×3) highlights ongoing benefits of localized visual cues.

Appendix A.4.1 for details). Humans always performed as well or better with the full image; performance never exceeded the patch-only condition. This robust pattern of context use contrasts with MLLMs showing negative PCRI, where patch-only input outperforms the full image: a deviation from human-like reasoning that signals shortcut exploitation or context brittleness. PCRI thus not only quantifies robustness, but also highlights departures from desirable, human-style contextual reasoning.

4.1.3 How does Task-Type modulate Context Sensitivity?

Captioning Tasks: Captioning (Figure 3) consistently show the strongest negative PCRI values, notably with InternVL2.5 (up to -0.49 at 2×2 and -0.75 at 3×3 for 2B). This suggests that image captioning inherently focuses on localized entities rather than global scene understanding, making these models particularly susceptible to background objects and noise.

Multiple-Choice QA (MCQ) Tasks: MCQ tasks (Figure 4) also exhibit negative PCRI scores, but with moderate sensitivity compared to captioning. Models such as Qwen2-VL and InternVL perform relatively better, supporting prior claims that MCQ tasks often leverage textual biases or selective visual attention mechanisms (Agrawal et al., 2016). However, models like Janus-Pro-7B and NVLM suffer significantly more, likely due to less sophisticated visual encoding strategies. Further details are in Appendix A.4.3.

Yes/No Classification Tasks: Yes/No tasks (Figure 5) exhibit the mildest PCRI scores, indicating that binary visual reasoning typically involves simpler or fewer visual cues, reducing the dependency on full-image context. Nevertheless, notable exceptions such as Pixtral-12B (-0.30 at 3×3) highlight significant variability and sensitivity, suggesting that model-specific factors such as visual encoder design affecting task robustness more than task complexity alone. Further details are in Appendix A.4.4 & Appendix A.4.2 for VQA tasks.

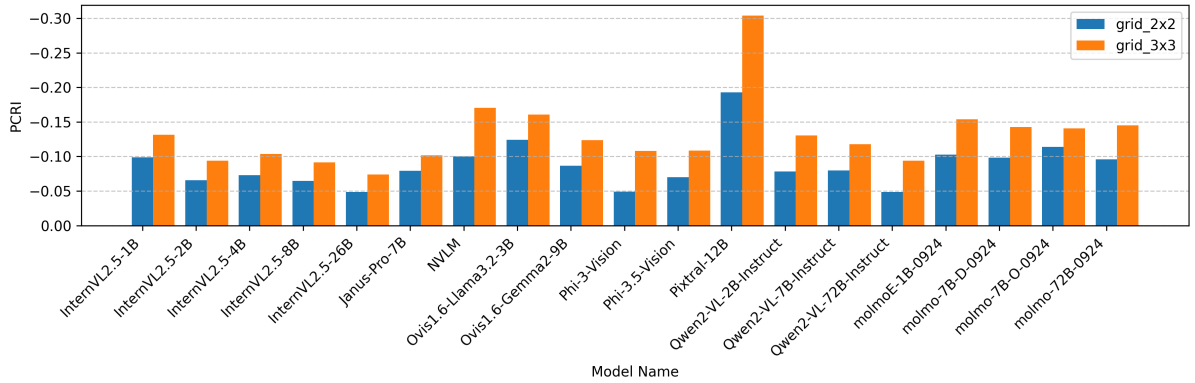

Figure 5: Avg. PCRI scores for Yes/No tasks across evaluated MLLMs. Most MLLMs show lower context sensitivity in binary decision-making scenarios compared to captioning tasks, but consistently improved performance with finer patch granularities (3×3) highlights ongoing benefits of localized visual cues.

4.1.4 Are all Models equally Context Sensitive?

Model-specific PCRI patterns reveal design trade-offs and robustness behaviors; see Appendix A.4.5 for details.

InternVL Models: InternVL variants exhibit varied context sensitivity across tasks. Strongly negative PCRI scores in captioning indicate that these models excel at fine-grained object recognition, yet struggle with holistic scene reasoning. Conversely, InternVL's relatively better context robustness on MCQ and Yes/No tasks likely stems from its dynamic resolution mechanisms and hierarchical attention layers, which facilitate effective selective encoding. Notably, the larger variant (26B) demonstrates superior resilience, validating the efficacy of hierarchical attention at larger scales.

Molmo Models: Molmo demonstrates relatively consistent PCRI values across tasks, highlighting strong robustness due to effective cross-modal alignment strategies. However, despite stable context robustness, its absolute performance is moderate in tasks demanding detailed visual reasoning (MCQ and captioning). This suggests that Molmo achieves robustness through generalized visual-textual alignment but at the cost of specialization for context-sensitive tasks.

Qwen2-VL Models: Qwen2-VL models show pronounced negative PCRI trends, particularly at lower scales, emphasizing their reliance on localized visual recognition strategies established through contrastive pre-training methods. The largest Qwen2-VL (72B) model notably achieves better robustness, likely benefiting from advanced multimodal rotary positional embeddings (M-RoPE) and resolution-adaptive encoding strategies, enhancing its global context integration capability.

4.1.5 How does image granularity ($n = 2, 3$) impact Model Robustness?

Increasing the granularity from 2×2 to 3×3 consistently amplifies negative PCRI scores across tasks (Figures 3,4,5), indicating that finer granularities further improve localized context performance relative to full-image contexts. Notably, larger models like 72B & 26B exhibit relatively lower PCRI magnitude shifts, suggesting that larger-scale hierarchical attention mechanisms provide enhanced robustness against visual context shifts. See Appendix A.4.6 for additional details and Appendix A.5 for qualitative examples.

4.2 Implications for Industry and Research

PCRI offers a rigorous, interpretable tool for evaluating MLLMs under real-world deployment constraints. Its key applications include:

- **Model Selection and Qualification:** PCRI provides a standardized criterion to identify & select models that maintain performance under variable or noisy visual conditions, supporting safer deployment in high-stakes domains (e.g., content moderation).

- **Model Design and Diagnosis:** By revealing context brittleness and patch sensitivity, PCRI pin points architectural weaknesses and guides targeted improvements, such as enhancing hierarchical attention, integrating retrieval-augmented modules, or optimizing cross-modal alignment.

- **Continuous Monitoring and Auditing:** In production, PCRI enables ongoing tracking of context robustness as data evolves, facilitating early detection of emerging vulnerabilities, crucial for regulatory compliance, user safety, and long-term reliability.

For practitioners, if PCRI is strongly negative, further investigation or model retraining is recommended; if near zero, the model can be trusted to generalize across visual contexts. See Appendix B for practical interpretation & real-world examples.

5 Conclusion

We introduce the Patch Context Robustness Index (PCRI), a scalable & interpretable score for systematically quantifying the sensitivity of MLLMs to changes in visual context granularity. Our evaluation, spanning 19 recent MLLMs & 15 diverse tasks, provides the first comprehensive study of context brittleness in the field.

Our analysis reveals that most MLLMs remain vulnerable to irrelevant or distracting context, with negative PCRI scores indicating performance degradation in the presence of full-scene information. In contrast, models such as InternVL2-26B and Qwen2VL-72B demonstrate superior context robustness across benchmarks, providing actionable choices for practitioners facing real-world noise and clutter. We also find substantial variation across task types, highlighting where global or local context is most essential.

PCRI enables direct comparison of model robustness, supporting both diagnostic evaluation & production deployment decisions. We demonstrate that prioritizing models with near-zero PCRI in our beta rollout led to measurably better user experience and reliability, even where task-level accuracy was matched.

By establishing a unified and extensible evaluation framework, our work lays the foundation for more robust, context-aware model selection and analysis in multimodal AI. Future directions include extending PCRI to sequential, video, and audio domains, enabling further advances in real-world robustness.

6 Limitations

While PCRI provides a robust, interpretable signal of context sensitivity in MLLMs, several limitations remain. First, the current analysis is restricted to English-language vision-language datasets; extending PCRI to multilingual and cross-cultural tasks is an important next step. Second, PCRI measures robustness at the granularity of patches, but does not explicitly account for dependencies across patches or sequential/temporal context, future work could address these aspects. Third, our framework focuses on task-level context sensitivity, rather than image property variations such as resolution or noise; integrating these factors remains an open challenge.

Finally, while we validate PCRI alignment with human reasoning across two task types, broader studies, including more diverse datasets and annotator pools, would further strengthen the generalizability of our findings. We encourage the community to adopt and extend PCRI as a diagnostic tool for developing and deploying trustworthy, context-robust multimodal models.

References

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*.

Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, Tao Sheng, Sujith Ravi, and Dan Roth. 2025a. Aligning llms for multilingual consistency in enterprise applications. *Preprint*, arXiv:2509.23659.

Amit Agarwal, Kulbhushan Pachauri, Iman Zadeh, and Jun Qian. 2024a. Techniques for graph data structure augmentation. US Patent 11,989,964.

Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, and Dong-Kyu Chae. 2025b. MVTamperBench: Evaluating robustness of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18804–18828, Vienna, Austria. Association for Computational Linguistics.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2024b. Synthetic document generation pipeline for training artificial intelligence models. US Patent App. 17/994,712.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2025c. FS-DAG: Few shot domain adapting graph networks for visually rich document understanding. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 100–114, Abu Dhabi, UAE. Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Hanning Chen, Yang Ni, Wenjun Huang, Yezi Liu, SungHeon Jeong, Fei Wen, Nathaniel Bastian, Hugo Latapie, and Mohsen Imani. 2024a. Vltp: Vision-language guided token pruning for task-oriented segmentation. *Preprint*, arXiv:2409.08464.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

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*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *Preprint*, arXiv:1504.00325.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024d. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024e. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Karan Dua, Puneet Mittal, Ranjeet Gupta, and Hitesh Laxmichand Patel. 2025. SpeechWeave: Diverse multilingual synthetic text & audio data generation pipeline for training text to speech models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 718–737, Vienna, Austria. Association for Computational Linguistics.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. 2024. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Preprint*, arXiv:1802.08218.

Chaoya Jiang, Haiyang Xu, Chenliang Li, Ming Yan, Wei Ye, Shikun Zhang, Bin Bi, and Songfang Huang. 2022. Trips: Efficient vision-and-language pre-training with text-relevant image patch selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4084–4096.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiongkuo Min, Xiaohong Liu, Weisi Lin, et al. 2025. R-bench: Are your large multimodal model robust to real-world corruptions? *IEEE Journal of Selected Topics in Signal Processing*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

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*.

Yiwei Ma, Zhibin Wang, Xiaoshuai Sun, Weihuang Lin, Qiang Zhou, Jiayi Ji, and Rongrong Ji. 2024. Inf-llava: Dual-perspective perception for high-resolution multimodal large language model. *arXiv preprint arXiv:2407.16198*.

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. *Preprint*, arXiv:2203.10244.

Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hard negative mining for domain-specific retrieval in enterprise systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1013–1026, Vienna, Austria. Association for Computational Linguistics.

Srikant Panda, Amit Agarwal, and Hitesh Laxmichand Patel. 2025a. Accesseval: Benchmarking disability bias in large language models. *Preprint*, arXiv:2509.22703.

Srikant Panda, Vishnu Hari, Kalpana Panda, Amit Agarwal, and Hitesh Laxmichand Patel. 2025b. Who's asking? investigating bias through the lens of disability framed queries in llms. *Preprint*, arXiv:2508.15831.

Srikant Panda, Hitesh Laxmichand Patel, Shahad Al-Khalifa, Amit Agarwal, Hend Al-Khalifa, and Sharefah Al-Ghamdi. 2025c. Daiq: Auditing demographic attribute inference from question in llms. *Preprint*, arXiv:2508.15830.

Hitesh Laxmichand Patel, Amit Agarwal, Arion Das, Bhargava Kumar, Srikant Panda, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, and Dong-Kyu Chae. 2025. Sweeval: Do llms really swear? a safety benchmark for testing limits for enterprise use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 558–582.

Hitesh Laxmichand Patel, Amit Agarwal, Bhargava Kumar, Karan Gupta, and Priyaranjan Pattnayak. 2024. Llm for barcodes: Generating diverse synthetic data for identity documents. *arXiv preprint arXiv:2411.14962*.

Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hybrid ai for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 215–229.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Bhargava Kumar, Amit Agarwal, Ishan Banerjee, Srikant Panda, and Tejaswini Kumar. 2024. Survey of large multimodal model datasets, application categories and taxonomy. *arXiv preprint arXiv:2412.17759*.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2024. Benchmarking robustness of multimodal image-text models under distribution shift. *Preprint*, arXiv:2212.08044.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

Jyotika Singh. 2021. Social media analysis using natural language processing techniques. In *Proceedings of the 20th Python in Science Conference*, SciPy, page 74–80. SciPy.

Jyotika Singh. 2022. pyaudioprocessing: Audio processing, feature extraction, and machine learning

modeling. In *Proceedings of the 21st Python in Science Conference*, SciPy, page 152–158. SciPy.

Jyotika Singh. 2023. *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*. Chapman and Hall/CRC.

Rahul Thapa, Kezhen Chen, Ian Covert, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2024. Dragonfly: Multi-resolution zoom-in encoding enhances vision-language models. *arXiv preprint arXiv:2406.00977*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

XAI-Org. 2024. Realworldqa dataset. `https://huggingface.co/datasets/xai-org/RealworldQA`. Accessed: 2024-03-15.

Wei Ye, Chaoya Jiang, Haiyang Xu, Chenhao Ye, Chenliang Li, Ming Yan, Shikun Zhang, Songhang Huang, and Fei Huang. 2024. Efficient vision-and-language pre-training with text-relevant image patch selection. *arXiv preprint arXiv:2403.07883*.

Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. 2024. Exploring perceptual limitation of multimodal large language models. *Preprint*, arXiv:2402.07384.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2023. Towards perceiving small visual details in zero-shot visual question answering with multimodal llms. *arXiv preprint arXiv:2310.16033*.

Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. 2025. Vcr: A task for pixel-level complex reasoning in vision language models via restoring occluded text. In *The Thirteenth International Conference on Learning Representations*.

Shiding Zhu, Wenhui Dong, Jun Song, Yanan Guo, and Bo Zheng. 2024. Hyvilm: Enhancing fine-grained recognition with a hybrid encoder for vision-language models. *arXiv preprint arXiv:2412.08378*.

A Appendix

A.1 Detailed Methodology: PCRI

Let $\mathcal{D} = \{(x^{(i)}, q^{(i)}, a^{(i)})\}_{i=1}^{N}$ denote a dataset of N image–query–answer triples, where $x^{(i)}$ is an image, $q^{(i)}$ is the associated query (e.g., question or prompt), and $a^{(i)}$ is the ground-truth answer.

Let $M(\cdot)$ be the evaluation metric for the task (e.g., accuracy, BLEU, F1), computed over $\mathcal{D}$ as per the standard benchmark protocol.

The model's performance on the dataset using the full image is:

$$P_{\text{whole}} = M\left(\{(x^{(i)}, q^{(i)}, a^{(i)})\}_{i=1}^{N}\right). \quad (2)$$

For each image $x^{(i)}$, let $\{x^{(i,j)}\}_{j=1}^{n^2}$ denote its $n \times n$ non-overlapping patches. The patch-based performance at granularity n is:

$$P_{\text{patch},n} = M\left(\{(x^{(i,j^*)}, q^{(i)}, a^{(i)})\}_{i=1}^{N}\right), \quad (3)$$

where $j^* = \arg\max_j s^{(i,j)}$ is the index of the patch with the highest model performance $s^{(i,j)}$ for instance i (with $s^{(i,j)}$ computed per the metric M).

The **Patch Context Robustness Index (PCRI)** at granularity n is:

$$\text{PCRI}_n = 1 - \frac{P_{\text{patch},n}}{P_{\text{whole}}}. \quad (4)$$

PCRI thus quantifies the (relative) performance drop or gain when the model is restricted to its best-performing local patch versus the full image context. Because P_{whole} and $P_{\text{patch},n}$ are computed with the same metric M over the same dataset, PCRI is *invariant* to the metric's scale and is directly comparable across tasks and datasets.

Interpretation.

- $\text{PCRI}_n \approx 0$: Model is robust—global and local context yield similar performance.
- $\text{PCRI}_n > 0$: Model requires global context; patch-only input reduces performance.
- $\text{PCRI}_n < 0$: Model is distracted by global context or benefits from local-only cues.

PCRI is undefined if $P_{\text{whole}} = 0$ for a given model/task pair, as division by zero is not meaningful. In such cases, the model cannot solve the task even with full context.

Rationale for Max Aggregation. For each instance, we select the patch j^* with the highest model performance $s^{(i,j)}$. This "max" aggregation captures whether the model can solve the task using *any* local patch. This is robust for both discrete and continuous metrics and highlights cases where a model is brittle to added context or reliant on specific local evidence. In contrast, mean or sum aggregation could mask context brittleness by averaging over patches that may be trivially correct or uninformative. The max operation thus yields a clearer, more actionable signal for model selection and robustness analysis.

Granularity sensitivity and compute. We ablate $n \in \{3, 4, 5\}$ on selected model and tasks (Table 2). Increasing n (finer granularity) can expose local solvability but increases evaluation cost quadratically (n^2 patch forward passes per image). For $n>3$, patches often become too small to capture meaningful semantics; PCRI's discriminative power saturates and may drop because the task becomes infeasible under heavy fragmentation. In practice, $n=2$ or $n=3$ provides a strong trade-off between informativeness and efficiency.

Dataset	$\text{PCRI}_{n=3}$	$\text{PCRI}_{n=4}$	$\text{PCRI}_{n=5}$
ChartQA_TEST	0.237	0.300	1.000
AMBER	-0.038	-0.030	0.550
BLINK	-0.516	-0.558	0.380

Table 2: PCRI values (Molmo-1B) at increasing patch granularity (n=3, 4, 5) on representative benchmarks. More negative ⇒ stronger global-context distraction (best patch ≥ full image). Very large positive values at n=5 indicate fragmentation-induced unsolvability rather than new trends, motivating our default of $n \in \{2, 3\}$.

A.2 Chance floors and Validity gate

Chance floors. For each dataset d, we define $C(d)$ as the task's random baseline: $1/|\mathcal{Y}|$ for balanced $|\mathcal{Y}|$-way classification; the class-prior baseline for imbalanced classification; and the documented random/shuffle baseline for retrieval/captioning metrics. When an official baseline is unavailable, we estimate $C(d)$ via random-shuffle following standard protocol.

How to set $C(d)$. *Classification:* $C(d) = 1/|\mathcal{Y}|$ if balanced; otherwise use the empirical class-prior baseline. *Retrieval:* for N candidates and one relevant item, $C_{\text{R@}K} \approx K/N$; if multiple relevants or

nonstandard pools, use the dataset's documented random baseline or Monte Carlo shuffle. *Captioning/QA metrics:* use the documented random or shuffle baseline provided by dataset authors.

Uncertainty and gate application. We estimate SE for P_{whole} via nonparametric bootstrap over evaluation examples (B=1000 by default) and apply the gate $P_{\text{whole}} \geq C(d) + \max\{\delta, 2\,\text{SE}\}$ with δ=0.01 on $[0, 1]$-scaled metrics.

Reporting policy. If a model–dataset pair fails the gate, do *not* compute/interpret PCRI; report only the underlying task metrics and mark PCRI as N/A. In our experiments, all model–dataset pairs satisfy the gate; we therefore interpret PCRI for all results.

A.3 Benchmarks & Models

We evaluate our approach across 15 widely-used vision-language benchmarks, covering a diverse range of tasks. These datasets were selected to represent a balanced mix of localized perception tasks (e.g. object recognition) and global contextual reasoning challenges (e.g., complex multi- choice question answering).

Our selection employs diverse realworld datasets that inherently contain a wide range of variations in image resolution, background complexity, and visual distortions. Consequently, PCRI metric has been rigorously tested across these naturally occurring variations, providing strong evidence of its robustness and practical relevance under realistic, heterogeneous visual conditions. Future work could complement these findings with controlled ablation studies to isolate the impact of each factor.

- **Visual Question Answering (VQA):** Benchmarks such as GQA (Lu et al., 2022), ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), and VizWiz (Gurari et al., 2018) are open-ended VQA tasks where MLLMs must generate responses without restricted answer choices. These benchmarks assess a model's ability to infer answers based on both localized and global scene information.

- **Multiple-Choice Question Answering (MCQ):** Benchmarks including BLINK (Fu et al., 2024b), RealWorldQA (XAI-Org, 2024), AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), and MMStar (Chen et al., 2024b) provide multiple answer choices, requiring MLLMs to differentiate between options and select the most accurate response. These tasks evaluate multimodal reasoning and answer disambiguation, offering insights into whether datasets support global contextual reasoning.

- **Yes/No Questions (Binary Classification):** Datasets such as POPE (Li et al., 2023), HallusionBench (Guan et al., 2024), AMBER (Wang et al., 2024a), and MME (Fu et al., 2024a) focus on binary (yes/no) questions. AMBER, in particular, tests the ability of the model to capture fine-grained spatial relationships, making it useful for evaluating whether a dataset enforces strict spatial comprehension.

- **Image Captioning and Semantic Understanding:** We include MS-COCO (COCO Captions) (Chen et al., 2015) to evaluate semantic understanding at the global level. Captioning tasks assess whether models can generate accurate textual descriptions based on an entire image rather than relying on isolated object-level information.

Models: We benchmark 19 state-of-the-art MLLMs, including InternVL (Chen et al., 2024d; Gao et al., 2024; Chen et al., 2024c,e), Janu, LlaVaOneVision (Li et al., 2024), Molmo (Deitke et al., 2024), Qwen2-VL (Wang et al., 2024b), Phi-3 Series (Abdin et al., 2024), Ovis (Lu et al., 2024) and Pixtral (Agrawal et al., 2024) across various model sizes.

A.4 Extended Results

This section provides detailed experimental results to supplement the main paper, including human-study, model and task-level PCRI breakdowns, granular ablations, and additional qualitative analyses. These extended results offer further evidence for the robustness patterns and context sensitivity described in the main text, supporting both reproducibility and deeper diagnostic insight for practitioners and researchers.

A.4.1 Human Study: Protocol and Production Validation

Protocol. To assess whether PCRI aligns with human context reasoning, we conducted a controlled study on two representative benchmarks: AI2D

Dataset	PCRI	Full Img. Perf.	Patch Perf.	Δ Perf.
AI2D	-0.23	96.7%	96.7%	0%
RealWorldQA	-0.29	98%	96%	-2%

Table 3: Human study accuracy (%) in patch vs. full image conditions, compared to model PCRI(InternVL-2.5-1B) scores.

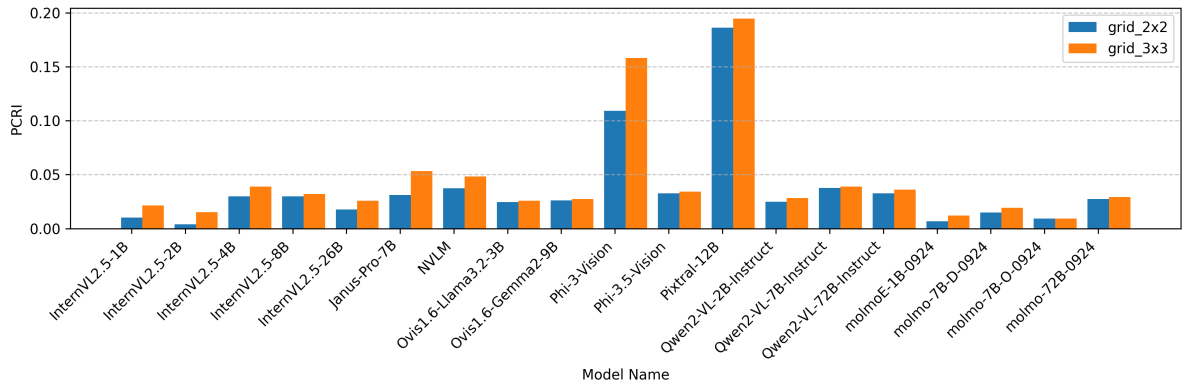

Figure 6: Avg. PCRI scores across different MLLMs for VQA tasks at patch granularities 2×2 and 3×3 Positive PCRI values indicate improved model performance with full-image contexts.

and RealWorldQA. For each, we randomly sampled approximately 20% of the data (AI2D: 300 samples; RealWorldQA: 80 samples). Three annotators (authors, blinded to model outputs and labels) independently answered each query in two settings: (1) with all the patches (one at a time), and (2) with the full image. Annotators strictly followed official task instructions and provided both answers and qualitative feedback for each condition.

Findings. Table 3 summarizes results. In all cases, human accuracy with the full image was equal to or higher than with any patch: for AI2D, performance was nearly identical in both settings; for RealWorldQA, accuracy increased slightly with the full image (from 96% patch-only to 98%). Annotators consistently reported higher confidence and less ambiguity when presented with the full image, reinforcing the value of global context for human reasoning.

By contrast, several MLLMs exhibited **negative PCRI** on these benchmarks, performing better on patches than on the full image, behavior never observed in humans. This suggests that such models are either overfitting to spurious cues in isolated patches or are distracted by irrelevant global information, leading to brittle and non-human-like reasoning. Qualitative annotator feedback further supports this diagnosis, with patch-only settings described as "ambiguous" or "lacking key context."

Production Deployment and User Feedback. To validate the practical impact of PCRI in real-world settings, we used PCRI to guide model selection for an internal MLLM-powered product during a closed beta launch (over 300 users, proprietary dataset). Models with near-zero PCRI were consistently preferred over negative-PCRI models, even when overall task accuracy was similar. User feedback highlighted greater reliability, consistency, and stability in outputs from models with higher context understanding, confirming that PCRI is predictive of real deployment success.

Conclusion. Our results confirm that negative PCRI is a non-human-like model behavior, while human reasoning mostly benefits from additional context. PCRI thus provides a valuable, interpretable signal for practitioners seeking robust, trustworthy MLLMs for production use.

A.4.2 PCRI on VQA Tasks

Visual Question Answering (VQA) tasks uniquely require integrating textual queries with visual details that may span broader contextual information across an image. Unlike MCQ or captioning tasks, where localized image patches consistently outperform full-image contexts (negative PCRI), our analysis reveals that VQA tasks exhibit a slight preference for full-image contexts, as indicated by uniformly small but positive PCRI scores (Figure 6).

1. Mild Positive PCRI: Preference for Broader Context Across the models evaluated, PCRI values for VQA tasks remain modestly positive (typically between 0.003 and 0.19), signifying slight but consistent benefits from broader image contexts compared to localized patches. This pattern diverges notably from the strong negative PCRI observed in Captioning and MCQ tasks, underscoring VQA's inherent requirement for integrating more comprehensive visual information and relational context rather than isolated visual details.

2. Model-specific Variability Significant variability exists across models. For instance, Pixtral-12B (PCRI=0.19 at 3×3) and Phi-3-Vision (PCRI=0.16 at 3×3) demonstrate the strongest preference for broader contexts.In contrast, InternVL2.5-2B and molmoE-1B-0924 show negligible PCRI scores (0.01), indicating minimal differentiation between localized and full-context settings.

3. Impact of Patch Granularity Increasing patch granularity from 2×2 to 3×3 generally results in moderately higher PCRI scores, indicating that models slightly prefer broader contexts (coarser granularity) over highly localized segments in VQA tasks. This suggests that global visual cues and contextual relationships become more salient and beneficial at lower granularities. However, the magnitude of these improvements remains moderate, implying that current MLLMs are already relatively effective at integrating visual information at coarser scales, and finer patches offer limited incremental advantages.

4. Contrasting VQA with MCQ and Captioning Tasks Unlike Captioning and MCQ tasks, where models consistently prefer localized patches (negative PCRI scores), the VQA task demonstrates an inherent need for broader visual context integration. This difference likely arises from the open-ended nature of VQA tasks, requiring more comprehensive visual reasoning and understanding of inter-object relationships and semantic context. This finding aligns with recent insights in the literature emphasizing the importance of global visual context for effective VQA reasoning (Jiang et al., 2022)

A.4.3 PCRI on MCQ Tasks

Multiple-Choice Question (MCQ) tasks require models to extract task-relevant information and choose a correct option to successfully solve the task. Figure 4 presents the PCRI across different models for 2×2 and 3×3 patch-based inputs, averaged over multiple MCQ datasets.

Across all models, PCRI remains negative, indicating that patch-based inputs consistently outperform full-image contexts. This suggests that global image representations potentially introduce unnecessary distractions or miss capturing necessary information during image encoding , diluting attention mechanisms and reducing task-specific accuracy.

We also observe that models generally perform better at $n = 3$ (more localized patches) compared to $n = 2$, reinforcing the hypothesis that MLLMs struggle to process global context effectively. Certain architectures, such as Molmo-1B, NVLM, and Janus-Pro, show more significant improvements in localized patch-based settings, implying that these models are particularly vulnerable to irrelevant background information. Conversely, larger-scale models like InternVL2.5-26B and Qwen2-VL-72B exhibit more stable PCRI, suggesting that increased capacity may improve context handling, though not entirely eliminate sensitivity to global context.

A.4.4 PCRI on Yes/No Question Answering Tasks

Yes/No question-answering tasks present a distinct challenge for multimodal models, as they often require binary reasoning over image content. Unlike open-ended VQA tasks, Y/N datasets tend to emphasize disambiguation of objects, attributes, or relationships within an image, making them an important benchmark for evaluating the impact of full-image versus localized context processing.

1. Localized Patches Improve Y/N Answering Across nearly all models, we observe consistently negative PCRI values, indicating that the models perform better when using localized patches rather than full-image input. This suggests that full-image representations introduce unnecessary context, leading to increased ambiguity in binary classification tasks.

2. Patch Size Influences Performance Gains When comparing 2×2 and 3×3 patches, the latter consistently yields lower PCRI values, meaning greater patch granularity improves performance. This supports the hypothesis that more focused image regions help models resolve Y/N questions by minimizing distractions.

3. Relationship Between Model Scale and Context Handling Larger models, such as Pixtral-12B, demonstrate a more stable PCRI, suggesting that scaling helps manage full-image context better. However, even for high-capacity models, localized patches still provide a performance boost, indicating that current attention mechanisms remain suboptimal for global reasoning in binary tasks.

These results further emphasize the need for adaptive attention filtering, where models dynamically adjust the level of contextual information they consider based on task requirements.

A.4.5 Model Level Analysis

Phi-3 Models: Phi-3 models consistently display moderate negative PCRI scores across tasks. Despite their compact size, Phi-3 architectures leverage efficient attention and optimized training strategies, mitigating severe performance degradation at higher granularities. However, their robustness still trails behind larger architectures like InternVL-26B and Qwen2-VL-72B, reflecting inherent capacity limitations for managing extensive visual context (Abdin et al., 2024).

Janus-Pro Models: Janus-Pro shows significant negative PCRI values across all tasks, notably severe in MCQ and captioning. This indicates substantial sensitivity to global context, attributable to its dual visual encoder approach, separately optimized for understanding and generation. Although Janus-Pro excels at specific multimodal generation tasks, its fragmented encoding strategy negatively impacts robustness in holistic scene comprehension, highlighting trade-offs in encoder specialization (Chen et al., 2025).

Ovis Models: Ovis variants present mixed PCRI scores, with smaller versions (Llama3.2-3B) heavily context-sensitive, while larger ones (Gemma2-9B) demonstrate improved robustness. The structured visual embedding strategy employed by Ovis provides a clear theoretical advantage for cross-modal embedding alignment, yet smaller variants still struggle with overwhelming contextual information due to limited embedding capacity. Larger Ovis models better leverage structured embeddings, balancing detailed visual perception with enhanced robustness (Lu et al., 2024).

Pixtral Models: Pixtral-12B exhibits notably high PCRI sensitivity, particularly in Yes/No and captioning tasks, suggesting challenges in effectively processing global contexts despite advanced ROPE-2D encoding strategies. This sensitivity highlights inherent trade-offs associated with high-resolution, multi-image reasoning, where detailed attention enhances local perception at the cost of global context integration (Agrawal et al., 2024).

A.4.6 Impact of Patch Granularity on PCRI Across Tasks

To further explore how varying visual granularities influence context robustness, we analyze the relative changes in PCRI scores when moving from 2×2 to finer-grained 3×3 patches across three primary task types: Multiple-choice QA (MCQ), Captioning, and Yes/No (Y/N) classification (Figure 7). This analysis complements the absolute PCRI evaluations (Figures 4, 3, 5) from our main results, highlighting important subtleties in the interaction between visual granularity and model performance. Performance across different models and datasets can be found in Table 5, 6, 7, and 8.

1. Interpreting Relative Drop in PCRI Figure 7 depicts the percentage drop in PCRI, measuring the magnitude of performance change when visual context granularity increases. Importantly, these relative changes must be interpreted in conjunction with absolute PCRI values to avoid misleading conclusions. A higher percentage drop may not necessarily indicate poor absolute robustness if initial PCRI magnitudes are small.

2. MCQ Tasks: High Relative Drops with Moderate Absolute Impact MCQ tasks exhibit the largest average relative PCRI drop (68.24%), suggesting that further granularity greatly enhances the advantage of localized patches over full images. For instance, InternVL2.5-26B shows an extreme relative drop (119.6%), yet its absolute PCRI remains relatively moderate (-0.036 at 2×2 to -0.079 at 3×3). Similarly, Qwen2-VL-72B-Instruct displays a 99.1% relative drop but maintains low absolute PCRI (-0.013 to -0.027). This highlights that hierarchical models exhibit substantial robustness to image and context granularity, with a slight improvement in performance on MCQ tasks as granularity increases.

3. Captioning Tasks: Moderate Relative Drops with Significant Absolute PCRI Captioning tasks have a smaller average relative PCRI drop (34.95%) compared to MCQ, but their absolute PCRI magnitudes are notably higher. For example,

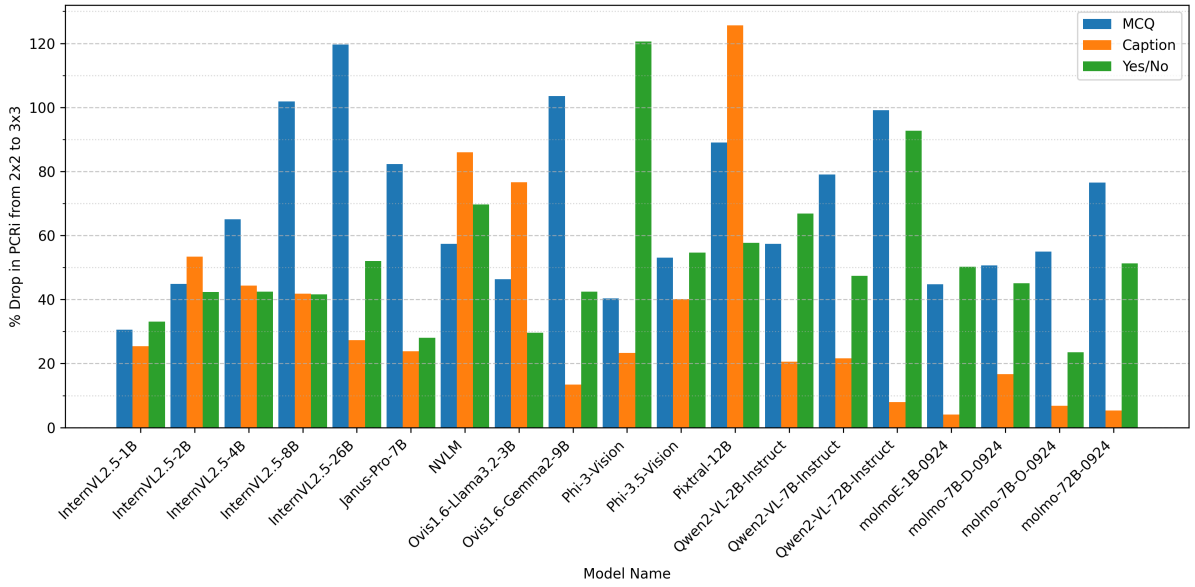

Figure 7: Percentage change in PCRI from 2×2 to 3×3 patches across MCQ, Caption, and Y/N datasets. A higher drop in PCRI suggests greater sensitivity to patch granularity, while a smaller drop indicates better stability in handling full-image contexts.

InternVL2.5-2B experiences a moderate relative change (53.4%) but displays significant absolute PCRI values (-0.492 at 2×2 to -0.755 at 3×3). Thus, captioning tasks consistently emphasize localized context, as previously noted, and even modest increases in granularity substantially amplify models' preference for localized patches, due to inherent task characteristics.

4. Yes/No Tasks: Balanced Relative and Absolute PCRI Values Yes/No classification tasks demonstrate intermediate behavior, with an average relative PCRI drop of 52.16% and moderate absolute PCRI magnitudes (typically below -0.15 at 3×3). Notably, some models like Phi-3-Vision (120.6% relative drop) and Qwen2-VL-72B (92.7% relative drop) exhibit significant sensitivity to granularity, indicating that while binary classification generally relies on simpler visual cues, specific architectural choices significantly influence robustness.

5. Architectural Implications and Context Sensitivity The variability observed underscores the complexity of interpreting PCRI in context. Models with sophisticated hierarchical attention (e.g., InternVL and Qwen2-VL larger models) tend to achieve strong absolute robustness, despite higher relative sensitivity to granularity. Conversely, smaller or simpler architectures experience pronounced absolute and relative degradation, highlighting critical vulnerabilities in their visual encoding and attention mechanisms.

A.5 Illustrative Examples of Context Granularity

To qualitatively illustrate how PCRI captures model sensitivity to context, Table 4 presents examples from RealWorldQA. For each question, we show the original image, its 2×2 and 3×3 patch splits, and highlight each patch: **green boundaries** indicate that the model answered correctly using only that patch, while **red boundaries** indicate incorrect predictions at that context level.

These examples highlight several key points:

- **Varied context requirements:** Some questions can be answered correctly from a single local patch (e.g., detection of a stop sign), while others require more global context or integration across regions (e.g., counting pedestrians).
- **PCRI as a diagnostic tool:** By analyzing which patches yield correct versus incorrect predictions, practitioners can diagnose local and global reasoning capabilities, as well as identify where context brittleness emerges.
- **Practical implications:** Such analysis informs targeted dataset curation, model development, and real-world deployment, by

Question	Original	2×2 Split	3×3 Split	Answerable with
Is the crosswalk sign active? Please answer directly with a single word or number.				Full image, 2×2 split (row 1, column 1), 3×3 split (row 2, column 1)
Is there a stop sign in this image? Please answer directly with a single word or number.				2×2 split (row 1, column 2), 3×3 split (row 2, column 2)
How many pedestrians are there? Please answer directly with a single word or number.				Full image

Table 4: Illustrative examples from the RealWorldQA dataset demonstrating how context granularity (full image vs. 2×2 and 3×3 splits) affects answerability and model accuracy. Green boundaries indicate correct answers at that patch/context; Red boundaries indicate incorrect predictions.

matching model selection and training to the true context requirements of end-user tasks.

Overall, these qualitative examples reinforce the value of PCRI as a fine-grained, interpretable measure of context robustness in vision-language models.

B Practical Implications and Usage Guide

The Patch Context Robustness Index (PCRI) offers a concrete, actionable tool for practitioners designing, auditing, and maintaining multimodal models for real-world applications. PCRI can be integrated throughout the MLLM lifecycle to:

- **Benchmark Model Robustness**: Rank candidate models on context sensitivity before deployment, ensuring chosen models remain reliable in environments with background clutter, occlusions, or incomplete views.
- **Monitor Deployed Systems**: Track PCRI over time to detect emerging vulnerabilities as data distributions shift, e.g., in dynamic environments or after retraining.
- **Auditing and Debugging**: Use PCRI to identify tasks or datasets where a model is brittle (negative or high PCRI), guiding further data collection or model refinement.
- **Compare Training Strategies**: Evaluate the effect of architectural choices or pretraining schemes on context robustness, providing a rigorous basis for model design.

B.1 Interpretation Guide

PCRI values provide actionable signals:

- **PCRI** ≈ 0: Model is robust; context granularity does not affect performance. Suitable for deployment in unpredictable visual conditions.
- **PCRI** > 0: Model requires global context (e.g., scene-level tasks); patch-only input insufficient.
- **PCRI** < 0: Model exploits local cues but fails globally, indicative of shortcut or brittle behavior; not human-like and poses deployment risk.
- **Undefined/large values**: Denominator is too small or zero; metric unreliable for that task/model combination.

Best practices: Use PCRI with other robustness metrics. Investigate negative PCRI to diagnose spurious correlations. Monitor PCRI regularly in production as model and data evolve.

B.2 Use-Cases

PCRI can be integrated at multiple stages of AI product development, deployment, and maintenance, offering value for robustness, reliability, and transparency in diverse settings:

- **Retail Product Search and E-Commerce:** Cluttered, dynamic shelves and varied camera angles can introduce substantial background

noise and distractors. PCRI enables teams to benchmark and select models that sustain high retrieval or classification accuracy despite irrelevant objects, reducing false positives and improving user trust in product recommendations, search and survellaince.

- **Assistive Technology for Accessibility:** For screen readers, object recognizers, and navigation aids used by visually impaired individuals, input images are often cropped, occluded, over-zoomed or partially visible. PCRI ensures selected models are robust in such scenarios, decreasing risk of missed cues or misleading outputs, and supporting safer user experiences in everyday environments.
- **Autonomous Vehicles and Robotics:** Changing backgrounds (construction, seasonal foliage, weather conditions, or dynamic obstacles) can degrade MLLM performance. By tracking PCRI over time, operators can identify when models become brittle to new environmental context, triggering targeted data augmentation, model retraining, or human-in-the-loop overrides before safety-critical failures.
- **Industrial Inspection and Quality Control:** Automated inspection systems (e.g., for manufacturing defects) must distinguish true faults from distracting background patterns or partial occlusions. PCRI supports the benchmarking of new models for robustness against such nuisance variation, guiding dataset augmentation and QA pipelines.
- **Content Moderation and Safety:** Social media and online platforms face adversarial attempts to evade detection via occlusion, cropping, or clutter. PCRI can flag models that are sensitive to such manipulations, helping design systems that maintain detection accuracy in the presence of adversarial context modification.
- **Medical Imaging and Diagnostics:** Context brittleness in radiology or pathology images can lead to missed findings or false alarms due to artifacts, cropping, or scanner noise. PCRI helps validate models on edge cases where only local detail is diagnostic, supporting higher reliability for clinical deployment and regulatory clearance.
- **Continuous Model Monitoring and Drift Detection:** In production, real-world data distributions evolve. Integrating PCRI into monitoring dashboards enables early detection of performance drift due to novel backgrounds or scene elements, supporting proactive retraining and minimizing negative user impact.
- **Model Regression Testing and Compliance Audits:** PCRI offers a standardized, quantitative metric for comparing successive model versions on context robustness, providing a clear "go/no-go" signal for deployment. Including PCRI in model cards or audit logs supports regulatory compliance and transparent documentation for stakeholders.
- **Benchmark and Dataset Curation:** PCRI can highlight underrepresented context challenges in existing benchmarks. Dataset designers can use PCRI analysis to guide new data collection—adding samples with cluttered, ambiguous, or challenging backgrounds to improve model generalization.
- **Internal Model Selection and Beta Testing:** As shown in our production beta launch (see Section A.4.1), models with higher (near-zero) PCRI delivered improved user feedback and reliability, even when overall accuracy was matched, highlighting PCRI's value for practitioner-facing decision-making.

B.3 Limitations and Caveats

- **Small denominators:** PCRI is unstable or undefined if few images are correct globally. Exclude such tasks or use bootstrapped intervals.
- **Patch granularity:** Excessively large n results in tiny, meaningless patches; use $n = 2$ or $n = 3$.
- **Metric agnosticism:** PCRI can use any base metric (accuracy, F1, etc.) as long as patch/whole scores are defined.
- **Independence:** PCRI ignores spatial dependencies between patches; future work may address this.

Model	AI2D		BLINK		MMMU		MMStar		RealWorldQA		ScienceQA	
	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$
InternVL2_5-1B	-0.17	-0.23	-0.05	-0.23	-0.21	-0.04	-0.22	-0.29	-0.05	-0.29	-0.31	-0.03
InternVL2_5-26B	-0.05	-0.03	-0.03	-0.09	-0.07	-0.06	-0.07	-0.16	-0.03	-0.13	-0.13	-0.05
InternVL2_5-2B	-0.09	-0.38	-0.01	-0.11	-0.20	-0.04	-0.13	-0.45	0.01	-0.23	-0.28	-0.02
InternVL2_5-4B	-0.08	-0.15	-0.04	-0.11	-0.17	-0.07	-0.10	-0.25	-0.04	-0.20	-0.25	-0.06
InternVL2_5-8B	-0.05	-0.01	-0.03	-0.09	-0.16	-0.07	-0.08	-0.10	-0.03	-0.17	-0.18	0.05
Janus-Pro-7B	-0.12	-0.50	-0.02	-0.19	-0.24	-0.01	-0.19	-0.70	0.02	-0.34	-0.62	-0.07
NVLM	-0.16	-0.47	-0.08	-0.13	-0.17	-0.05	-0.22	-0.70	-0.15	-0.24	-0.29	-0.10
Ovis1.6-Gemma2-9B	-0.04	-0.20	-0.02	-0.04	-0.12	-0.04	-0.08	-0.33	-0.08	-0.15	-0.18	-0.01
Ovis1.6-Llama3.2-3B	-0.13	-0.45	-0.10	-0.23	-0.17	-0.04	-0.17	-0.70	-0.14	-0.32	-0.24	-0.07
Phi-3-Vision	-0.09	-0.24	-0.08	-0.24	-0.26	-0.04	-0.12	-0.35	-0.11	-0.33	-0.33	0.02
Phi-3.5-Vision	-0.10	-0.04	-0.06	-0.27	-0.31	0.02	-0.13	-0.10	-0.09	-0.37	-0.45	-0.01
Pixtral-12B	-0.18	-0.12	-0.12	-0.17	-0.14	-0.13	-0.27	-0.37	-0.28	-0.38	-0.23	-0.17
Qwen2-VL-2B-Instruct	-0.11	-0.20	-0.04	-0.26	-0.17	-0.07	-0.16	-0.37	-0.04	-0.39	-0.27	-0.12
molmo-72B-0924	-0.12	-0.31	-0.02	-0.03	-0.09	-0.02	-0.16	-0.43	-0.07	-0.11	-0.17	-0.12
molmo-7B-D-0924	-0.12	-0.33	-0.07	-0.13	-0.10	0.00	-0.15	-0.41	-0.12	-0.23	-0.20	-0.03
molmo-7B-O-0924	-0.13	-0.27	-0.08	-0.18	-0.11	-0.03	-0.17	-0.40	-0.12	-0.29	-0.20	-0.04
molmoE-1B-0924	-0.22	-0.38	-0.17	-0.30	-0.22	-0.04	-0.32	-0.52	-0.27	-0.46	-0.31	-0.08

Table 5: PCRI MCQ scores for different models across datasets.

Model	AMBER		HallusionBench		MME		POPE	
	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$
InternVL2_5-1B	-0.01	-0.27	-0.06	-0.05	-0.03	-0.35	-0.10	-0.06
InternVL2_5-26B	-0.01	-0.10	-0.02	-0.06	-0.03	-0.15	-0.04	-0.07
InternVL2_5-2B	-0.06	-0.22	-0.05	-0.05	0.04	-0.27	-0.08	-0.06
InternVL2_5-4B	-0.01	-0.17	-0.05	-0.05	-0.03	-0.26	-0.06	-0.07
InternVL2_5-8B	-0.02	-0.16	-0.03	-0.05	-0.04	-0.22	-0.05	-0.05
Janus-Pro-7B	-0.01	-0.22	-0.05	-0.05	0.00	-0.28	-0.07	-0.06
NVLM	-0.02	-0.27	-0.07	-0.04	-0.04	-0.49	-0.09	-0.05
Ovis1.6-Gemma2-9B	-0.03	-0.19	-0.08	-0.05	-0.05	-0.26	-0.12	-0.06
Ovis1.6-Llama3.2-3B	-0.04	-0.29	-0.10	-0.07	-0.06	-0.37	-0.13	-0.08
Phi-3-Vision	-0.03	-0.09	-0.09	-0.05	-0.01	-0.16	-0.20	-0.06
Phi-3.5-Vision	-0.04	-0.14	-0.11	-0.08	0.02	-0.22	-0.14	-0.09
Pixtral-12B	-0.17	-0.36	-0.06	-0.18	-0.21	-0.65	-0.09	-0.26
Qwen2-VL-2B-Instruct	-0.07	-0.22	-0.11	-0.06	-0.02	-0.35	-0.12	-0.07
Qwen2-VL-72B-Instruct	-0.01	-0.10	-0.03	-0.05	-0.04	-0.22	-0.05	-0.07
Qwen2-VL-7B-Instruct	-0.02	-0.19	-0.06	-0.05	-0.05	-0.27	-0.09	-0.06
molmo-72B-0924	-0.04	-0.23	-0.06	-0.05	-0.06	-0.35	-0.11	-0.06
molmo-7B-D-0924	-0.02	-0.20	-0.12	-0.06	-0.04	-0.33	-0.14	-0.06
molmo-7B-O-0924	-0.04	-0.26	-0.11	-0.05	-0.05	-0.34	-0.10	-0.06
molmoE-1B-0924	-0.02	-0.22	-0.13	-0.04	-0.04	-0.35	-0.17	-0.05

Table 6: PCRI YN scores for different models across additional datasets.

Model	ChartQA		GQA TestDev Balanced		TextVQA VAL		VizWiz	
	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$	$PCRI_{n=2}$	$PCRI_{n=3}$
InternVL2_5-1B	0.202	-0.207	0.063	-0.099	0.190	-0.266	0.112	-0.122
InternVL2_5-26B	0.264	-0.159	0.080	-0.115	0.266	-0.215	0.121	-0.149
InternVL2_5-2B	0.260	-0.210	0.078	-0.143	0.263	-0.276	0.128	-0.175
InternVL2_5-4B	0.286	-0.175	0.086	-0.077	0.257	-0.221	0.123	-0.083
InternVL2_5-8B	0.302	-0.142	0.076	-0.116	0.282	-0.179	0.127	-0.103
Janus-Pro-7B	-0.035	0.115	-0.095	-1.231	-0.173	0.320	-0.052	-2.231
NVLM	0.193	-0.164	0.062	-0.242	0.169	-0.223	0.108	-0.248
Ovis1.6-Gemma2-9B	0.251	-0.262	0.090	0.131	0.189	-0.322	0.155	0.211
Ovis1.6-Llama3.2-3B	0.166	-0.251	0.070	1.000	0.105	-0.333	0.135	1.000
Phi-3-Vision	0.198	-0.160	0.070	-0.545	0.145	-0.212	0.107	-1.273
Phi-3.5-Vision	0.080	-0.190	-0.025	0.265	0.026	-0.259	-0.008	0.184
Pixtral-12B	0.228	-0.212	0.102	0.626	0.178	-0.258	0.153	0.667
Qwen2-VL-2B-Instruct	0.243	-0.189	0.075	-0.030	0.190	-0.265	0.119	0.006
Qwen2-VL-72B-Instruct	0.231	-0.137	0.095	-0.059	0.226	-0.179	0.145	-0.048
Qwen2-VL-7B-Instruct	0.288	-0.162	0.099	-0.075	0.257	-0.217	0.152	-0.076
molmo-72B-0924	0.217	-0.199	0.111	-0.020	0.167	-0.255	0.156	0.010
molmo-7B-D-0924	0.211	-0.229	0.113	-0.036	0.163	-0.283	0.145	-0.020
molmo-7B-O-0924	0.199	-0.224	0.097	-0.036	0.139	-0.296	0.136	-0.016
molmoE-1B-0924	0.198	-0.235	0.093	-0.083	0.137	-0.310	0.136	-0.092

Table 7: PCRI VQA scores for different models across additional datasets.

Model	**COCO VAL**	
	PCRI$_{n=2}$	PCRI$_{n=3}$
InternVL2_5-1B	-0.2429	-0.3047
InternVL2_5-26B	-0.4921	-0.7547
InternVL2_5-2B	-0.1009	-0.1285
InternVL2_5-4B	-0.2773	-0.4005
InternVL2_5-8B	-0.3941	-0.5588
Janus-Pro-7B	-0.1302	-0.1613
NVLM	-0.1665	-0.3097
Ovis1.6-Gemma2-9B	-0.1447	-0.1042
Ovis1.6-Llama3.2-3B	-0.0932	-0.0646
Qwen2-VL-2B-Instruct	-0.0231	-0.0079
Qwen2-VL-72B-Instruct	-0.0354	0.0082
Qwen2-VL-7B-Instruct	-0.0492	-0.0598
molmo-72B-0924	-0.1286	-0.1354
molmo-7B-D-0924	-0.1714	-0.1599
molmo-7B-O-0924	-0.0534	-0.0570
molmoE-1B-0924	-0.1304	-0.1357

Table 8: PCRI Caption scores for different models on COCO dataset.