SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

SPARC-Rad: A Multimodal Benchmark Dataset and Evaluation Pipeline for Spatial and Anatomical Reasoning in Radiology Vision-Language Models

Satvik Tripathi, University of Pennsylvania; Dania Daye, MD, PhD; Tessa Cook, MD, PhD, CIIP, FSIIM

Introduction/Background

Vision-language models (VLMs) are increasingly evaluated for medical image interpretation, yet conventional benchmarks primarily emphasize disease recognition, report generation, or general visual question answering. These tasks may not adequately assess spatial reasoning, including anatomical localization, laterality, orientation, and relationships between structures. We developed the Spatial Perception and Anatomical Reasoning in Clinical Radiology benchmark (SPARC-Rad) to directly evaluate anatomically grounded spatial reasoning across contemporary VLMs.

Methods/Intervention

SPARC-Rad comprised 300 manually curated image-question pairs derived from radiologic studies available through The Cancer Imaging Archive. The benchmark included CT (n=98), MRI (n=88), and radiography (n=114), spanning abdomen (n=71), chest (n=64), breast (n=57), neuroimaging (n=54), and musculoskeletal imaging (n=54). Questions assessed anatomical identification, localization, laterality, regional recognition, inter-structure relationships, projection-based reasoning, and cross-sectional orientation. Fourteen general-purpose, open or cloud-hosted, and medical-domain VLMs were evaluated using standardized prompting. Responses were scored using an LLM-assisted framework followed by human validation. Overall, modality-stratified and anatomy-stratified accuracy were calculated with 95% confidence intervals. Pairwise model differences were assessed using chi-square and matched McNemar tests.

Results/Outcome

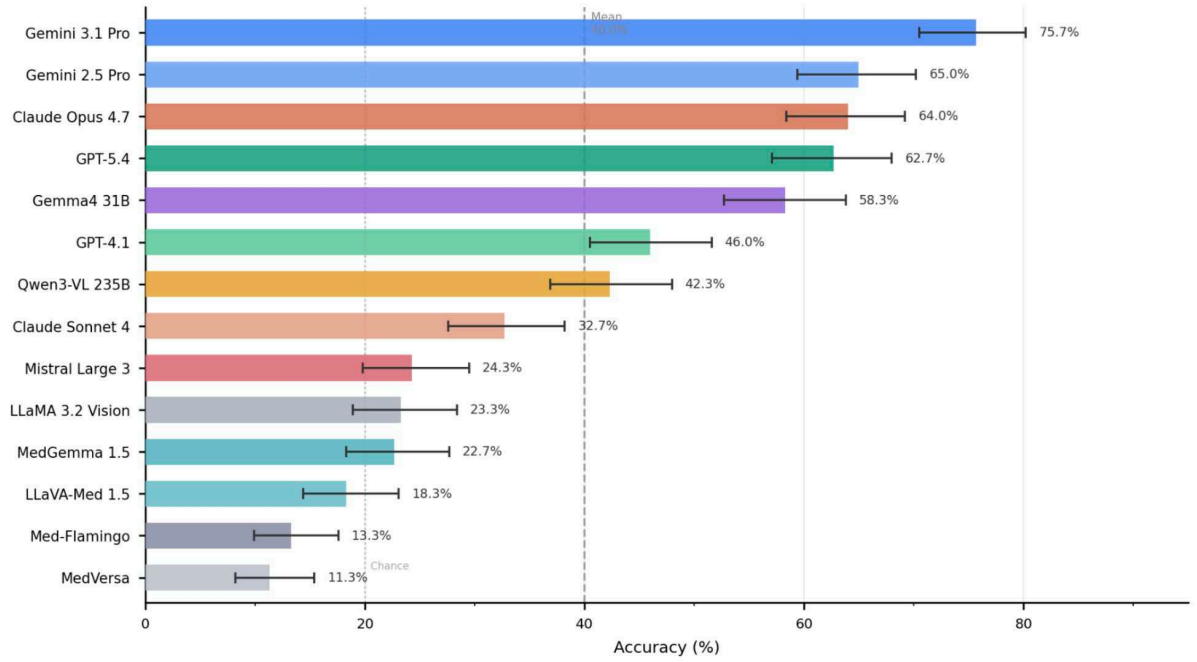
Overall accuracy varied from 11.3% to 75.7% across models, with a mean accuracy of 40.0% and a standard deviation of 21.1 percentage points. Gemini-3.1-Pro-Preview achieved the highest overall accuracy at 75.7% (227/300; 95% CI, 70.5%–80.2%), followed by Gemini-2.5-Pro-Preview (65.0%), Claude-Opus-4.7 (64.0%), and GPT-5.4 (62.7%). Performance was modality-dependent: Gemini-3.1-Pro-Preview performed best on CT (83.7%) and MRI (79.5%), whereas GPT-5.4 performed best on radiography (69.3%). Across anatomical categories, chest imaging had the lowest mean accuracy (35.4%), while abdomen had the highest (43.4%). Clinically relevant errors included laterality confusion, incorrect localization, regional misidentification, relational reasoning failures, and difficulty interpreting projection anatomy. Significant differences were identified in 76 of 91 chi-square comparisons and 79 of 91 matched McNemar comparisons.

Conclusion

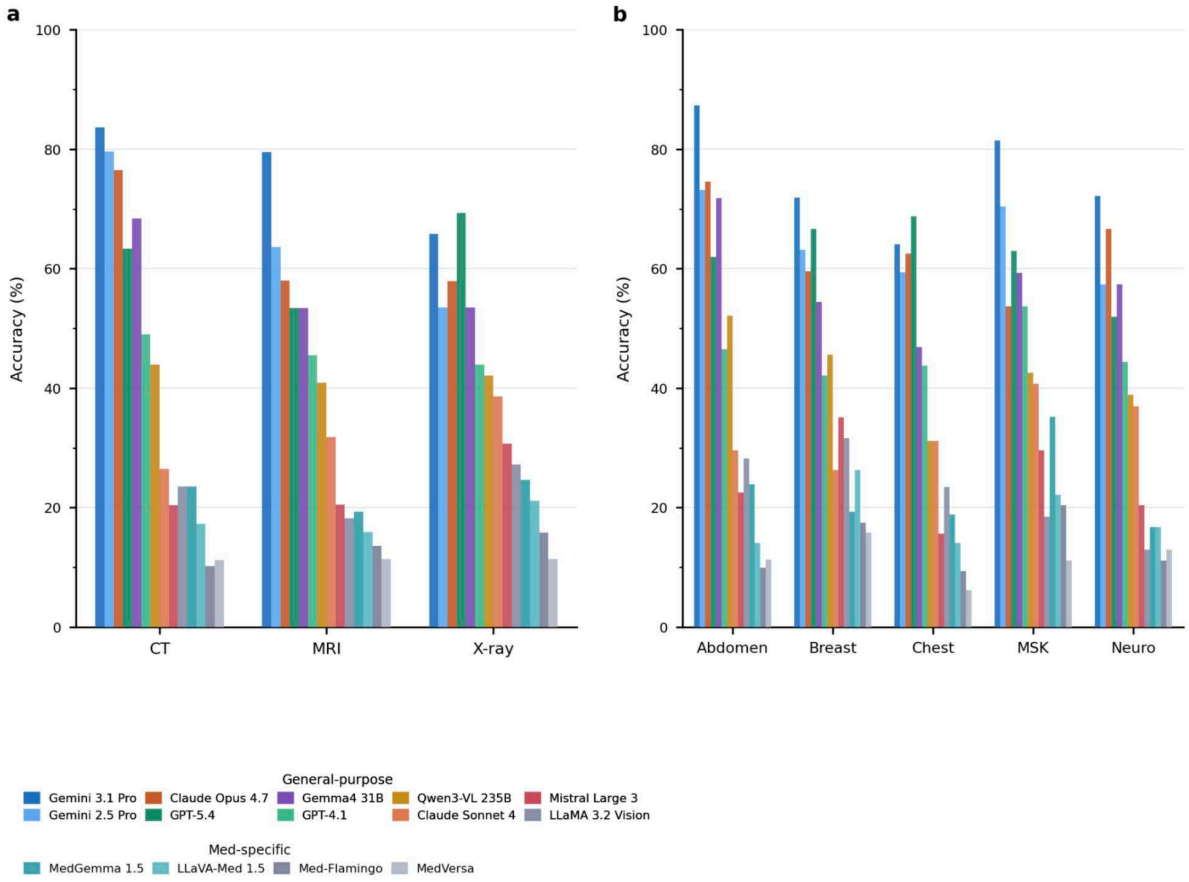
Current VLMs demonstrate substantial and model-specific limitations in spatial and anatomical reasoning. Therefore, direct evaluation of spatial and anatomical relationships should complement conventional diagnostic benchmarks before clinical deployment.

Statement of Impact

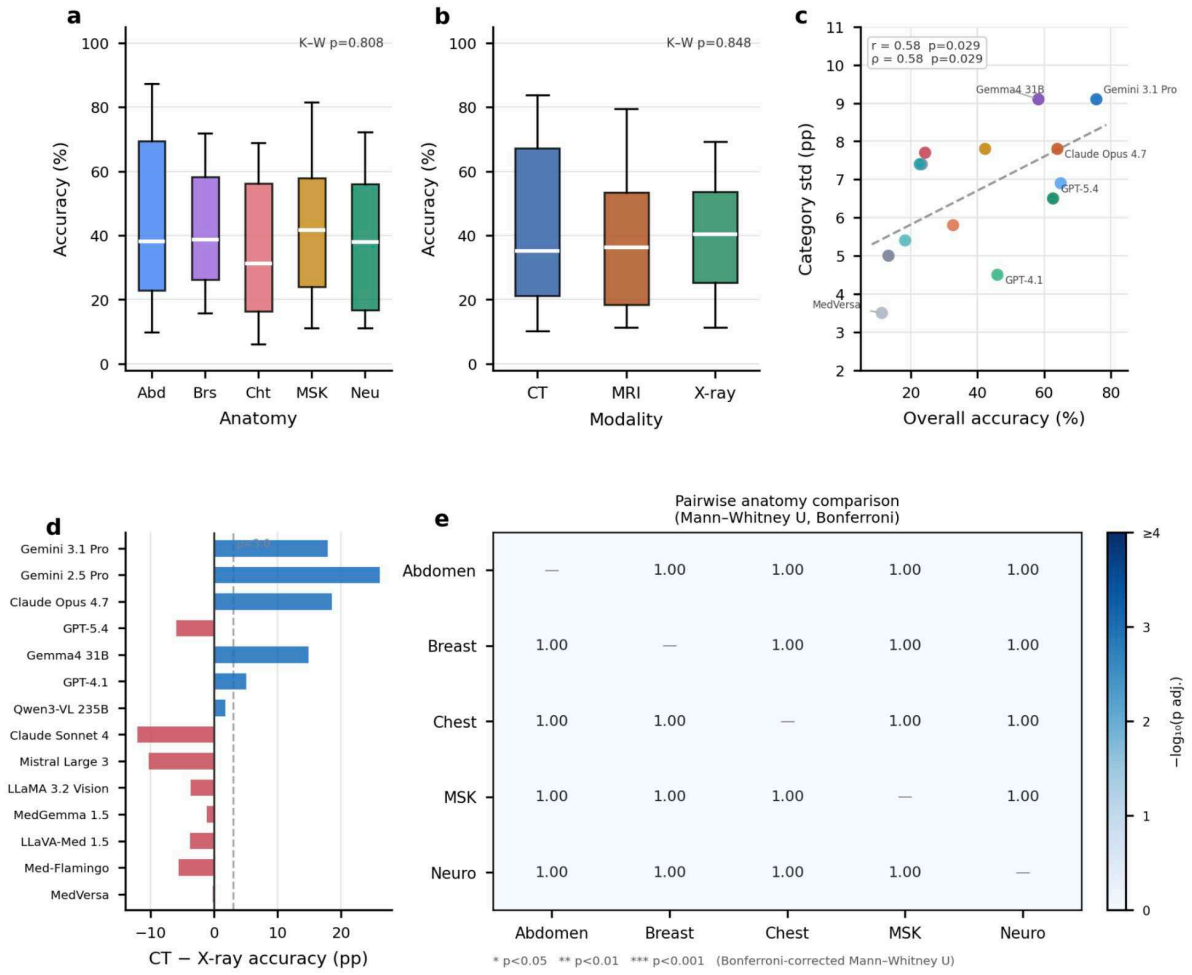
SPARC-Rad identifies a clinically important gap between recognizing medical images and understanding their spatial anatomy. Incorporating spatially grounded evaluation into model development and validation may help prevent clinically consequential localization and laterality errors and establish more realistic standards for medical AI readiness.



Overall model performance on SPARC-RAD



Performance stratified by modality and anatomy



Statistical analysis of the model performance across modality and anatomy

Keywords

VLM; Dataset; Evaluation