

Multimodal Vision-Language Models for Explainable Clinical Document Understanding and Medical Image-Based Decision Support

Ivan Hansen

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.

helloivan@uc.edu

Leihao Xu

Department of Electrical Engineering and Computer Science, University of Missouri,

Columbia, MO, USA.

helloleihao@missouri.edu

Neeraj R. Bansal

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

contactneeraj@unh.edu

Abstract

Modern clinical systems produce heterogeneous records that combine narrative notes, structured tables, scanned forms, medical images, and operational data. The integration of vision and language in artificial intelligence has opened new possibilities for clinical document understanding and image-based decision support, but it also raises substantial questions about explainability, governance, fairness, and deployment. This paper examines multimodal vision-language models from an interdisciplinary systems perspective, focusing on how architectural choices interact with clinical accountability. It analyzes the alignment of visual and textual representations, the role of layout-aware document understanding, the grounding of medical image interpretation in contextual evidence, and the structural trade-offs between model complexity and transparency. The discussion emphasizes that explainability should not be treated as a post hoc interface but must be embedded in architecture, evaluation, institutional workflows, and regulatory practice. The paper further considers data governance, bias, sustainability, and policy implications, arguing that clinical adoption depends on robust evidence tracing, adversarial auditing, and adaptive oversight rather than predictive performance alone. The analysis connects recent technical advances to the operational realities of health systems and concludes with forward-looking governance recommendations for accountable multimodal clinical systems.

Keywords

multimodal learning, clinical document understanding, medical imaging, explainable artificial intelligence, health system governance.

1. Introduction

Modern health care is characterized by an overwhelming mixture of clinical text, structured laboratory results, scanned documents, medical images, and administrative records. The cognitive work of synthesizing these modalities remains a persistent bottleneck in diagnosis, treatment planning, quality measurement, and clinical research. For many years, machine

learning addressed language and vision through largely separate pipelines. This separation conflicts with the reality of clinical reasoning, which is inherently multimodal. The development of attention-based transformer architectures changed sequence modeling and enabled more flexible representation learning [1]. Bidirectional pretraining subsequently improved contextual language understanding across many clinical text tasks [2]. Vision transformers extended transformer-based reasoning to image recognition by treating image patches as sequences [3]. Contrastive multimodal pretraining brought language and vision into a shared representational space [4], while bootstrapped language-image pretraining refined this alignment through generative tasks [5]. Instruction-tuned visual language models expanded these capabilities toward conversational and task-focused reasoning [6]. In medicine, foundation models have begun to target generalist capabilities across modalities and institutions [7], and large language models have been shown to encode substantial clinical knowledge [8]. These advances suggest that multimodal models may soon support clinical document interpretation and medical image-based decision support in routine practice. However, clinical deployment requires more than benchmark accuracy. It requires systems that can explain their outputs, integrate with regulated workflows, and remain robust under distributional shift. This article examines these requirements from a systems perspective, with particular attention to the structural tensions among model complexity, explainability, fairness, and sustainability.

A core challenge is that clinical documents are not ordinary free text. They combine narrative statements, coded fields, tables, stamps, signatures, and references to prior episodes, often rendered as scanned pages or PDFs with irregular layouts. Traditional optical character recognition and rule-based extraction systems degrade when confronted with such variability. Vision-language models offer a potential solution by jointly processing visual layout and linguistic content. The capacity to understand tables, forms, and free text in a unified framework is particularly relevant for clinical notes, discharge summaries, and laboratory reports. Recent work has focused on table recognition using vision language models, underscoring how layout-aware visual understanding can transform document digitization pipelines [9]. However, clinical deployment demands that such recognition be tightly coupled to downstream verification, uncertainty estimation, and audit trails, rather than treated as a standalone extraction task.

The promise of multimodal vision-language models in clinical domains extends beyond document digitization to include medical image interpretation, report generation, and clinical decision support. A central theme in this article is that explainability cannot be reduced to attention maps or saliency overlays. It must be understood as a property of the entire sociotechnical system, including model architecture, training data, evaluation protocols, deployment infrastructure, and institutional governance. The following sections systematically address these dimensions, moving from architectural considerations to fairness, sustainability, and regulatory implications.

2. Multimodal Architectures for Clinical Document Understanding

Clinical document understanding is fundamentally different from general domain document parsing because clinical records are governed by strict confidentiality rules, professional terminology, and legally significant content. Architectures designed for this task must reconcile two competing demands. The first demand is representational power sufficient to capture fine-grained visual layout such as table structures, checkboxes, stamps, and

handwritten annotations, while the second demand is interpretability sufficient to support clinical accountability. Transformer-based vision encoders partition images into patches and learn self-attention across those patches, allowing the model to integrate global layout information with local text regions [3]. When combined with text encoders, these models can align visual regions with corresponding words, enabling tasks such as form understanding, key-value extraction, and table recognition. The architectural trade-off is that deeper cross-modal fusion improves accuracy but obscures the pathway from input tokens to output labels, making error tracing more difficult.

One promising direction is the use of detection and grounding mechanisms within multimodal models. By explicitly predicting bounding boxes or segmentation masks for relevant regions, models can provide spatial explanations that link extracted fields to their source coordinates. This form of visual grounding is especially valuable in clinical settings because it allows a human auditor to verify whether a model has read the correct portion of a scanned document. However, grounding alone is not sufficient for explainability. The model may correctly identify the region but fail to interpret its semantic content due to domain shift or rare abbreviations. Therefore, architectures must include mechanisms for confidence calibration and abstention, such as learned uncertainty heads or selective prediction layers. These mechanisms are not post hoc additions but must be designed into the training objective and decoding procedure.

Another important architectural consideration is the choice between general-purpose foundation models and domain-adapted models. General-purpose models such as CLIP and BLIP demonstrate strong zero-shot performance on many visual language tasks [4][5], yet they often lack the specialized vocabulary and layout sensitivity required for clinical documents. Domain adaptation through continued pretraining on de-identified clinical corpora and medical forms can improve performance, but it also introduces new governance risks related to data provenance and potential memorization of protected health information. From a systems perspective, the optimal architecture depends not only on benchmark accuracy but also on the ease with which errors can be audited, the availability of labeled data for fine-tuning, and the computational budget of the deploying institution.

3. Visual Grounding and Explainability in Medical Imaging

Medical imaging presents a different set of challenges for multimodal vision-language models. Unlike document understanding, where the primary goal is to extract and link structured information, medical imaging requires interpretation of anatomical structures, pathological findings, and temporal changes. Vision-language models can generate radiology reports from chest radiographs, brain magnetic resonance imaging, or computed tomography scans. The visual grounding problem becomes one of aligning image regions with clinically meaningful concepts such as lesions, effusions, or fractures. Such grounding is essential for explainability because it enables the model to indicate where a finding appears in the image and to cite supporting visual evidence for its textual conclusion.

The structural trade-off in medical imaging is between generative language outputs and reference-based retrieval. Generative systems can produce fluent and coherent reports, but they may hallucinate findings that are not present in the image. Retrieval-based systems, by contrast, select relevant sentences from an existing report corpus and ground them in retrieved similar images. Retrieval-based approaches often provide stronger evidence traces because each generated statement can be linked to a source report and image. However, they may be less flexible when encountering rare or complex presentations that do not match existing

examples. Hybrid architectures that combine generative and retrieval components aim to balance fluency with verifiability, but they introduce additional complexity in training and inference.

Explainability in medical imaging also requires attention to counterfactual reasoning and contrastive explanations. A model that states "pneumothorax present" should be able to answer why it considers the finding present and what changes in the image would alter its conclusion. Contrastive explanations can be generated by perturbing image regions, text descriptions, or model activations and observing the resulting changes in output. While such methods can illuminate model behavior, they are computationally intensive and may not be stable across different perturbation strategies. Clinical systems should therefore treat contrastive explanations as one component of a broader explanation interface that also includes uncertainty scores, differential diagnosis lists, and pointers to relevant training exemplars.

A critical governance issue is the distinction between model explanation and clinical explanation. A model explanation describes why an algorithm produced a particular output; a clinical explanation justifies a decision in terms understandable to clinicians and patients. The two are often conflated in technical literature. Vision-language models can support clinical explanation by generating natural language justifications, but those justifications must be verified against medical knowledge and institutional guidelines. In this regard, the integration of external knowledge graphs or curated ontologies can improve the factual consistency of generated explanations, although it also raises questions about how to synchronize knowledge sources with evolving clinical practice.

4. Governance, Fairness, and Regulatory Implications

The deployment of multimodal vision-language models in health care occurs within a dense regulatory environment that includes data protection laws, medical device regulations, and professional liability frameworks. In the United States, the Food and Drug Administration has begun to develop guidance for adaptive artificial intelligence and machine learning-enabled medical devices, recognizing that models may continue to learn after deployment. In the European Union, the Artificial Intelligence Act classifies certain medical AI applications as high risk, imposing requirements for risk management, data governance, transparency, and human oversight. These regulatory developments create structural incentives for model developers to prioritize explainability and robustness over raw performance.

Fairness is a particular concern because multimodal models can inherit biases from both visual and textual training data. Clinical documents and medical images are generated under conditions of unequal access to health care, and they may reflect disparities in diagnostic testing, treatment, and documentation quality. A vision-language model trained on such data may learn to associate certain demographic groups with specific disease labels or to under-detect findings in images from underrepresented populations. Fairness audits must therefore examine not only aggregate performance but also subgroup-specific error rates, calibration, and the distribution of generated explanations. Explainability can support fairness by making visible the visual or textual features that most influenced a model output, allowing auditors to identify spurious correlations.

A significant governance challenge is the need for longitudinal monitoring and post-market surveillance. Clinical models operate in nonstationary environments where disease prevalence, imaging equipment, documentation practices, and clinical guidelines change over time. A

model that is explainable at deployment may become less interpretable as it is fine-tuned on new data or as its input distribution shifts. Institutions must implement monitoring systems that track performance degradation, explanation drift, and changes in fairness metrics. Such monitoring should be integrated with incident reporting and model versioning so that decisions can be audited retrospectively. Without such infrastructure, explanations offer little protection against silent failures.

5. Deployment, Infrastructure, and Sustainability

Clinical deployment of multimodal vision-language models requires substantial computational and data infrastructure. Training large vision-language models consumes significant energy and computational resources, and inference in real-time clinical environments introduces additional latency and cost constraints. The sustainability of these systems is rarely addressed in academic evaluations, which tend to focus on accuracy metrics rather than total cost of ownership. Health systems must consider whether the environmental footprint of a large generalist model is justified relative to a smaller domain-specific model with similar clinical performance. Model compression, distillation, and quantization can reduce inference costs, but they may alter the explanation behavior of the model in unpredictable ways.

The architectural choice between centralized and federated learning also has implications for governance and sustainability. Centralized training simplifies model development and evaluation, but it requires transferring sensitive patient data to a central repository, which may be prohibited or undesirable. Federated learning allows models to be trained across multiple institutions without sharing raw data, preserving privacy while potentially improving generalizability. However, federated systems introduce new complexities in model aggregation, versioning, and explainability. An explanation generated by a federated model may be difficult to attribute to specific training data, complicating audits and accountability. Hybrid approaches that share only model gradients or synthetic features represent a middle ground, but they require robust differential privacy guarantees to prevent re-identification.

Sustainability also encompasses the maintainability of explanation interfaces and the training of human users. Clinicians will not adopt a system whose explanations require specialized machine learning knowledge to interpret. Explanation interfaces must be co-designed with end users, including physicians, nurses, medical coders, and patients. This multidisciplinary design process adds time and cost to deployment, but it is essential for building trust and avoiding automation bias. Furthermore, institutions must invest in ongoing training and support so that clinicians can critically evaluate model outputs rather than accepting them passively. Without such human factors engineering, even the most technically sophisticated explainability methods will fail in practice.

6. Evaluation, Robustness, and Adversarial Auditing

Evaluation of multimodal clinical systems must extend beyond traditional accuracy measures to include explainability quality, robustness under distributional shift, and resilience to adversarial manipulation. Explainability quality is difficult to quantify because different stakeholders have different expectations. Clinicians may value concise, evidence-linked justifications, while regulators may require complete audit trails and uncertainty quantification. Patients may prioritize plain-language explanations that respect their health literacy and emotional state. A multimodal evaluation framework should therefore include multiple explanation criteria, such as faithfulness, completeness, stability, and actionability.

Faithfulness refers to the degree to which an explanation accurately reflects the model's internal reasoning, while completeness captures whether all relevant factors are mentioned.

Robustness under distributional shift is particularly important in clinical settings because image acquisition protocols, document templates, and coding standards vary across institutions and over time. A vision-language model that performs well on data from one hospital system may fail on data from another due to differences in image contrast, scanner calibration, or documentation style. Domain adaptation techniques can mitigate these failures, but they often reduce explainability because the adapted model may develop new attention patterns or feature representations that are not well understood. Adversarial auditing should include deliberate perturbation of inputs, such as adding noise to images, reordering document sections, or simulating rare disease presentations, to assess whether explanations remain consistent and informative.

Adversarial attacks also pose a threat to clinical systems. Malicious actors could modify medical images or documents in ways that are imperceptible to human clinicians but cause the model to produce incorrect outputs. Explainability can play a defensive role by making anomalous visual or textual features more visible, but the attacker may also target the explanation mechanism itself. For example, an adversary might craft an input that produces a plausible but misleading explanation while hiding the true source of error. Robust explanation methods must therefore be evaluated under adversarial threat models, and clinical systems should include independent verification layers that do not rely solely on the model's own explanation output.

Finally, the evaluation of explainability must be tied to clinical outcomes and institutional learning. It is not enough to demonstrate that a model can generate plausible justifications in a retrospective study. Health systems should conduct prospective evaluations that measure whether the use of explanations changes clinical decision quality, reduces diagnostic errors, or improves patient trust. Such evaluations are methodologically challenging because clinical outcomes are influenced by many confounding factors, and blinding is often impossible. Nevertheless, without outcome-oriented evidence, explainability remains an academic abstraction rather than a practical mechanism for accountability.

7. Conclusion

Multimodal vision-language models represent a significant advance in the capacity to interpret clinical documents and medical images within unified computational frameworks. Their ability to align visual regions with linguistic concepts supports tasks ranging from table extraction in scanned reports to grounded radiology report generation. Yet the transition from research prototype to clinical infrastructure requires a fundamental shift in how these systems are designed, evaluated, and governed. Explainability must be treated as a first-class system property, embedded in architecture, training objectives, monitoring pipelines, and user interfaces, rather than as a post hoc visualization layer. The structural trade-offs among model complexity, transparency, robustness, and sustainability demand careful navigation by health systems that operate under resource constraints and regulatory obligations.

A recurring theme of this analysis is that technical explainability and clinical accountability are not synonymous. A model can produce a high-quality visual grounding and a fluent natural language justification while still failing to support a defensible clinical decision. Accountability requires institutional mechanisms for auditing model behavior, tracking data provenance, monitoring distributional drift, and responding to errors with clear lines of

responsibility. These mechanisms are as important as the model architecture itself. The integration of vision-language models into clinical workflows should therefore be guided by interdisciplinary teams that include not only machine learning engineers and clinicians but also ethicists, patient representatives, legal experts, and operations researchers.

Looking forward, the most pressing research challenges involve developing explanation methods that are faithful under distributional shift, fairness audits that account for multimodal sources of bias, and evaluation protocols that measure the impact of explanations on real clinical decisions. The use of table recognition in vision language models and the continued evolution of augmented reality for surgical guidance illustrate the breadth of clinical tasks that may benefit from multimodal alignment [9][18]. Future work should also explore how regulatory frameworks can adapt to models that continue to learn from institutional data while preserving patient privacy and explanation stability. By addressing these challenges systematically, the field can move toward clinical systems that are not only accurate but also trustworthy, sustainable, and accountable.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.
5. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning*, 162, 12888–12900.
6. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.
7. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265.
8. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180.

9. Liu, Y., & Zhu, Y. (2025, May). Enhancing Table Recognition Using Vision Language Models (VLM). In 2025 IEEE 5th International Conference on Computer Communication and Artificial Intelligence (CCAI) (pp. 143-147). IEEE.
10. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
12. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.
13. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.
14. Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Shaban, M., Shady, M., Williams, M., Joo, B., Noor, Z., Valieris, R., Zollner, M., Topol, G. J., Kather, J. N., & Mahmood, F. (2023). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 41(2), 292–307.
15. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 119.
16. European Commission. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
17. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.
18. Yang, Y., Necker, F., Leuze, C., Chen, M., Finegersh, A., Lee, J., ... & Baik, F. M. (2026). Markerless augmented reality registration for surgical guidance: a multi-anatomy clinical accuracy study. *International Journal of Computer Assisted Radiology and Surgery*, 1-8.
19. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
20. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.