Reinforcement Learning and Vision-Language Models for Intelligent Clinical Report Generation and Medical Image Understanding

Sawyer Woods
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
swoods@buffalo.edu

Tejeas A. Senght
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
singhtejas@unh.edu

Leonard Hamilton
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
leonard.work@missouri.edu

Abstract

The automated generation of clinical reports from medical images is a grand challenge at the intersection of computer vision, natural language generation, and healthcare delivery. Recent advances in vision-language models (VLMs) have enabled remarkable zero-shot and few-shot image understanding, while reinforcement learning (RL) offers a principled framework for optimizing sequence-level clinical objectives that are non-differentiable and align with diagnostic quality. This paper presents a system-level interdisciplinary analysis of the integration of RL and VLMs for clinical report generation and medical image understanding. We examine architectural paradigms that combine large-scale pretrained multimodal representations with reward-driven fine-tuning tailored to radiological correctness, factual consistency, and clinical utility. The discussion moves beyond algorithm design to consider structural trade-offs involving data governance, infrastructure federation, model robustness, fairness across populations, and regulatory compliance. We analyze how RL-guided relation extraction and graph-structured reasoning can sharpen the alignment between visual findings and textual narratives, while confronting challenges such as hidden stratification, demographic bias, and the risk of fluent but clinically inaccurate generations. The paper further addresses deployment sustainability, energy consumption of foundation models, continuous monitoring in live hospital environments, and the evolving policy landscape defined by FDA action plans and the EU AI Act. By synthesizing technical architectures with sociotechnical governance, we argue that the next generation of intelligent clinical reporting systems must be designed as accountable, transparent, and continuously validated socio-technical infrastructures. The analysis identifies open problems and advocates for interdisciplinary collaboration across machine learning, medicine, ethics, and law to realize safe and equitable clinical AI.

Keywords

reinforcement learning, vision-language models, clinical report generation, medical imaging, fairness, governance, deployment, sociotechnical systems.

1. Introduction

The rapid digitization of healthcare has produced vast repositories of medical images paired with free-text radiology reports, creating an unprecedented opportunity for machine learning to assist clinical decision-making. Deep learning systems for medical image classification have demonstrated expert-level performance in tasks such as pneumonia detection from chest radiographs [1], while large-scale labeled datasets like CheXpert [2] and the MIMIC-CXR database of chest images with associated reports [3] have fueled research into automated interpretation. Moving beyond classification toward the automatic generation of coherent, clinically meaningful reports represents a far more complex challenge that requires reasoning about multiple findings, their anatomical locations, disease severities, and the nuanced language of radiology. Early work on medical report generation drew upon image captioning architectures enhanced with attention mechanisms [4], later incorporating memory-driven transformers to capture long-range dependencies [5] and knowledge distillation to leverage prior clinical knowledge [6]. These approaches, however, typically optimize token-level likelihood objectives that do not directly capture the clinical quality, factual completeness, or diagnostic utility of the generated text. Reinforcement learning emerged as a compelling alternative, with self-critical sequence training demonstrating that non-differentiable metrics such as BLEU or CIDEr can be directly optimized for image captioning [7], and this principle has since been extended to medical contexts. Simultaneously, the rise of vision-language models pretrained on web-scale image-text pairs has introduced powerful multimodal representations that can align visual concepts with natural language with remarkable fidelity, opening new pathways for domain-specific adaptation. Yet realizing the potential of these technologies in clinical settings demands far more than algorithmic innovation. The design, deployment, and governance of systems that couple RL with VLMs for clinical report generation raise structural questions about data infrastructure, fairness, robustness, regulatory compliance, and human-machine collaboration. This paper offers an interdisciplinary systems-level examination of these intersecting themes, positioning the integration of reinforcement learning and vision-language models not merely as a machine learning problem but as a sociotechnical infrastructure challenge.

2. Background and Paradigm Shifts in Multimodal Clinical Intelligence

The transition from unimodal image classifiers to report generation systems parallels a broader paradigm shift in artificial intelligence toward foundation models that jointly model vision and language. The release of CLIP [8] demonstrated that contrastive pretraining on 400 million image-text pairs could produce representations with strong zero-shot transfer, radically simplifying downstream tasks by aligning visual and linguistic modalities in a shared embedding space. Adapting such models to the medical domain, however, requires careful handling of domain-specific terminology, the high cost of annotation errors, and the profound consequences of misinterpretation. MedCLIP [9] addressed this by introducing contrastive learning from unpaired medical images and text, reducing the need for exact image-report correspondences and enabling the use of large-scale clinical data warehouses where pairing is noisy or incomplete. These methods produce rich multimodal encoders that can serve as the backbone for report generation, yet they remain largely unsupervised with respect to the structured clinical relations that radiologists encode in their reports, such as the anatomical location of an opacity, its connection to a specific disease, and differential diagnoses. Recent work on multi-modal relation extraction uses reinforcement learning-guided graph diffusion to jointly model visual and textual entities, dynamically building

structured representations that connect image regions to clinical concepts [10]. By framing the extraction of clinical relations as a sequential decision process under a reward that encourages factual consistency, this line of research bridges the gap between perceptual alignment and structured reasoning, an essential step for generating reports that are not only fluent but medically precise.

In parallel, large-scale biomedical vision-language models exemplified by BiomedCLIP [11] have been pretrained on millions of biomedical image-text pairs from the scientific literature, capturing a far richer vocabulary of diseases, anatomical structures, and imaging modalities than general-domain models. The emergence of zero-shot diagnostic systems such as CheXzero [12] further underscores the power of VLM-based approaches by demonstrating chest X-ray interpretation without explicit disease labels, using only comparison with free-text descriptions of findings. While these achievements mark significant progress in image understanding, generating full narrative reports requires bridging the gap between concept recognition and coherent clinical discourse. Direct reinforcement learning optimization for radiology report generation has been explored through architectures such as the Reinforced Transformer [13], which fine-tunes a sequence-to-sequence model using policy gradient methods to maximize metrics like BLEU and clinical accuracy proxies. The synergy between large pretrained VLMs and RL-based sequence refinement thus emerges as a natural design pattern: the VLM provides a semantically grounded initialization, while RL injects the ability to optimize for holistic, task-specific clinical criteria that are fundamentally non-decomposable at the token level.

3. System Architecture and Integration of Reinforcement Learning with Vision-Language Models

Designing a clinical report generation system that tightly integrates RL and VLMs entails a series of architectural decisions that have far-reaching implications for performance, safety, and maintainability. A canonical architecture consists of a pretrained VLM encoder that ingests medical images and optional prior textual context, a cross-modal fusion module that aligns visual features with language embeddings, and a decoder that generates the report token by token. The decoder is typically bootstrapped with supervised fine-tuning on paired image-report corpora to ensure basic fluency and domain vocabulary acquisition. In the subsequent RL phase, the model generates full reports which are evaluated by a clinically informed reward function that may combine lexical metrics, factual consistency scores derived from information extraction models, and, where available, structured labels from radiology information systems. Policy gradient methods such as REINFORCE or proximal policy optimization adjust the decoder parameters to increase the probability of reports that earn higher reward, thereby shifting the model's behavior from mimicking average human reports to producing those that satisfy explicitly articulated clinical objectives.

This architecture introduces critical trade-offs that must be understood from a systems engineering perspective. The choice of reward function is perhaps the most consequential design element, as it encodes institutional priorities, risk tolerances, and clinical standards. A reward that heavily weights syntactic fluency may produce articulate but clinically vacuous text, whereas overemphasizing the presence of specific findings can yield stilted prose that disrupts the clinician's reading flow. Multifacted rewards that penalize false positives and false negatives in finding mentions, while rewarding lexical diversity and adherence to structured reporting templates, require careful calibration and validation on diverse patient populations. Furthermore, the interaction between the VLM backbone and the RL fine-tuning

introduces the risk of catastrophic forgetting of the general knowledge encoded in the pretrained weights. Continual learning strategies, regularization terms, and mixed-objective training regimes that blend supervised next-token prediction with policy gradient updates are necessary to preserve a balance between domain adaptability and the retention of broad multimodal reasoning capabilities.

The computational infrastructure required to support such a system extends from large-scale pretraining clusters to inference engines deployed within hospital network boundaries. Pretraining a biomedical VLM from scratch may consume thousands of GPU-hours and raises questions about the environmental sustainability of such efforts when considered across an ecosystem of competing healthcare AI developers. Fine-tuning with RL, while typically less resource-intensive, introduces additional complexity in the form of parallel rollout collection, reward model serving, and careful management of multiple model checkpoints for reproducible evaluation. In a production clinical environment, the latency constraints are stringent, as reports may need to be generated in near real-time to support emergency radiology workflows. This compels architects to explore model distillation, quantization, and hardware-aware inference optimization without sacrificing the nuanced clinical reasoning that the RL-VLM integration is designed to achieve.

4. Data Governance, Infrastructure, and Multimodal Integration Challenges

Clinical data that fuel training and evaluation of RL-VLM systems are subject to some of the most stringent privacy and security regulations worldwide. Scans and reports contain protected health information, and even de-identified datasets remain vulnerable to re-identification when linked with external data sources. As a result, the data infrastructure must support federated learning paradigms that allow model training across multiple hospital sites without centralizing sensitive records. Federated RL, in which local policy gradients are computed on institutional data and securely aggregated, introduces additional layers of system complexity related to non-identical data distributions across sites, variable annotation quality, and the need for synchronized reward definitions that maintain clinical consistency across institutions. The governance of such distributed pipelines necessitates technical mechanisms such as differential privacy guarantees, secure multi-party computation, and robust consent management frameworks that respect patient autonomy while enabling meaningful model improvement.

Equally important is the semantic integration of heterogeneous data modalities that extend beyond single images. Comprehensive clinical report generation benefits from longitudinal patient histories, prior reports, laboratory values, and clinical notes, all of which must be fused with the current imaging study. Extending RL-VLM architectures to incorporate this multimodal context requires designing attention mechanisms and memory modules that can selectively attend to relevant historical information while suppressing spurious correlations. The VLMs that underpin image understanding must also contend with the diversity of imaging protocols, equipment manufacturers, and patient positioning, demanding robust domain adaptation and calibration procedures that can detect and flag out-of-distribution inputs before they corrupt the generated text. Implementing such safeguards demands an enterprise-grade software stack that orchestrates data loading, preprocessing, model inference, log collection, and continuous monitoring within a highly regulated environment where software updates may themselves require regulatory re-approval.

5. Robustness, Fairness, and Clinical Validation

The deployment of RL-VLM report generation systems in real clinical workflows brings ethical and technical challenges regarding their robustness and fairness across diverse patient groups. Landmark studies have revealed that deep learning models for chest X-ray classification can exhibit significant performance disparities across racial, gender, and socioeconomic subgroups, a phenomenon rigorously analyzed in the CheXclusion investigation [14]. Such disparities, if propagated into generated reports, could systematically under-report critical findings for underserved populations, thereby reinforcing existing healthcare inequalities. The problem is compounded by hidden stratification, where models perform well on average but fail on clinically meaningful subgroups defined by rare comorbidities, atypical presentations, or demographic intersections [15]. The RL component, while capable of optimizing for a composite reward, can exploit unintended shortcuts if the reward function does not explicitly penalize subgroup-specific failures. Designing reward functions that are fairness-aware, incorporating constraints on disparities across protected attributes, requires careful legal and ethical deliberation, as group fairness metrics such as equalized odds can conflict with individual clinical accuracy.

Robustness must also be assessed under distribution shifts induced by changes in imaging equipment, patient populations, or disease prevalence over time. A report generation system that was trained on pre-pandemic data and then exposed to COVID-19 pneumonia cases faces a fundamental challenge of generating clinically sound text for novel radiological patterns. Continuous post-deployment monitoring using prospectively collected data, combined with outlier detection and human-in-the-loop fallback mechanisms, becomes an integral component of the system architecture. Without such mechanisms, the system risks silently generating misleading reports that erode clinician trust. Ethical frameworks for machine learning in healthcare underscore the necessity of embedding fairness, accountability, and transparency into the entire lifecycle, from design to decommissioning [16]. The healthcare AI community has long recognized that big data alone does not guarantee clinically useful models without rigorous prospective validation [17], and that high-performance medicine depends on seamless integration of human expertise with machine intelligence [18]. Comprehensive guides to deep learning in healthcare stress the importance of multidisciplinary design teams, inclusive data curation, and external validation on diverse populations [19], a message reinforced by systematic reviews confirming that while deep learning matches healthcare professionals in many imaging tasks, the evidence base still lacks sufficient real-world clinical trials [20].

6. Deployment, Sustainability, and Continuous Learning

Transitioning an RL-VLM report generation prototype into a clinically deployed system requires navigating a complex landscape of operational constraints, sustainability considerations, and evolving maintenance requirements. The deployment environment typically encompasses hospital picture archiving and communication systems, electronic health records, and radiology information systems, each with its own data formats, access control policies, and latency profiles. The report generation engine must interoperate with these legacy systems through standardized interfaces such as DICOM and HL7 FHIR, while maintaining strict audit trails for every generated report to enable retrospective analysis and legal accountability. The computational footprint of running a large VLM plus an RL-decoder on every chest radiograph can be substantial, raising both cost and energy consumption concerns. In settings where dozens of hospitals are served by a centralized cloud inference service, the cumulative carbon impact invites scrutiny from sustainability-conscious

healthcare organizations and regulators. Mitigations include adopting smaller distilled models that retain core clinical competencies, leveraging sparsity and conditional computation, and deploying on-premises edge accelerators that keep data local and reduce network transfer overhead.

Sustainability also encompasses the human and organizational dimensions. The introduction of automated report generation alters the workflow of radiologists, who must shift from primary authorship to verification and editing of machine-generated drafts. This redefinition of professional roles demands careful change management, user interface design that clearly differentiates machine outputs from human-authored text, and continuous feedback loops that allow radiologists to flag errors and contribute corrections that can be used to refine the model through ongoing RL cycles. Such a continuous learning setup, however, introduces risks of feedback-induced data drift and overfitting to the correction patterns of a small group of clinicians. Governance protocols must therefore define when and how model updates are authorized, which data is eligible for retraining, and how to maintain comparability of performance metrics across model versions. The notion of algorithmic auditing frameworks becomes highly relevant, as internal and external audits can identify performance regressions, emergent biases, and unintended behavioral shifts that a purely automated monitoring system might overlook.

7. Policy Implications and Regulatory Landscape

Intelligent clinical report generation sits at the intersection of multiple regulatory domains, including medical device regulation, data protection law, and emerging artificial intelligence governance frameworks. Language models that produce clinical text pose unique regulatory challenges because of their open-ended output space, which complicates the exhaustive pre-market testing typical of traditional software as a medical device. The U.S. Food and Drug Administration has outlined an action plan for AI and machine learning-based software as a medical device that acknowledges the need for a total product lifecycle approach, in which manufacturers continuously monitor and improve algorithms after deployment while adhering to predetermined change control plans [24]. An RL-VLM system that continues to learn from radiologist edits in the field would need to be situated within such a regulatory framework, with clearly specified boundaries on permissible autonomous adaptation. The European Union's Artificial Intelligence Act classifies certain medical AI applications as high-risk, subjecting them to requirements for transparency, human oversight, robustness, and data governance [25]. For report generation, this likely entails providing clinicians with an explanation of how the report was derived, the ability to override or modify any section, and ongoing documentation of the model's performance across demographic strata. The interaction between RL's adaptive nature and the static certification paradigm represents a fundamental tension that regulators and developers must jointly resolve.

Beyond compliance, policy levers can shape the trajectory of the field toward more equitable outcomes. Procurement rules in public health systems can mandate the use of open datasets, community-developed benchmarks that measure fairness and robustness, and the publication of model cards that transparently communicate intended uses, limitations, and performance across subgroups. The broader societal discourse around large language models has raised concerns about their tendency to generate plausible yet incorrect information, a phenomenon that in the clinical domain could directly harm patients if reports contain fabricated findings or omit critical abnormalities. Mitigating such risks requires not only technical safeguards but also clear liability frameworks that delineate responsibility among developers, deploying

institutions, and supervising physicians. Multinational harmonization of these frameworks is essential as healthcare AI solutions are increasingly marketed across borders, yet divergent cultural attitudes toward automation and medical authority complicate the creation of uniform standards.

8. Conclusion

The integration of reinforcement learning and vision-language models for clinical report generation and medical image understanding represents a frontier where advances in machine learning must be tempered and guided by the realities of healthcare delivery. This paper has presented a systemic analysis that moves from architectural design choices—the selection of VLM backbones, the formulation of rewards, and the orchestration of RL fine-tuning—to the broader considerations of data governance, fairness, robustness, deployment sustainability, and regulatory alignment. We argued that the most significant risks in clinical report generation stem not from a lack of technical capability but from failures in sociotechnical integration: mismatched incentives between model optimization and clinical outcomes, unexamined biases encoded in training data, and an absence of robust continuous validation pipelines. The technical strength of reinforcement learning in optimizing non-differentiable clinical metrics must be harnessed within frameworks that enforce transparency, fairness constraints, and human oversight. The required reference highlights how RL-guided graph diffusion can improve structured relation extraction from multimodal data, illustrating one pathway toward more factually grounded generation. Yet even the most sophisticated relation extraction and narrative generation algorithms will fail to deliver equitable benefit unless accompanied by deliberate investments in diverse and representative data ecosystems, inclusive model auditing, and adaptive regulatory structures. Looking forward, interdisciplinary collaboration must intensify to develop standardized evaluation protocols that go beyond lexical similarity to measure clinical actionability, to design federated learning infrastructures that respect patient privacy while enabling collective model improvement, and to create governance mechanisms that keep pace with the rapid evolution of generative AI capabilities. The vision of an intelligent clinical reporting assistant that augments rather than replaces human expertise is within reach, but its realization demands a commitment to building not just better algorithms but trustworthy, accountable, and sustainable health intelligence infrastructures.

References

1. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

2. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 590–597.

3. Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1), 317.

4. Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2577–2586).

5. Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1439–1449).

6. Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021). Exploring and distilling posterior and prior knowledge for medical report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13767–13776).

7. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7008–7024).

8. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763).

9. Wang, Y., Li, H., Xiao, J., Wang, X., & Li, T. (2022). MedCLIP: Contrastive learning from unpaired medical images and text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8391–8406).

10. Yang, R., & Gupta, R. (2025, January). Enhancing Multi-Modal Relation Extraction with Reinforcement Learning Guided Graph Diffusion Framework. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 978-988).

11. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., ... & Poon, H. (2024). BiomedCLIP: a multimodal biomedical foundation model pretrained from diverse sources. arXiv preprint arXiv:2405.08796.

12. Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., & Rajpurkar, P. (2022). CheXzero: Chest X-ray diagnosis with zero-shot learning. Nature Biomedical Engineering, 6(12), 1386–1397.

13. Xue, Y., Huang, L., Wang, J., Zhang, J., & Cao, Y. (2021). Reinforced transformer for medical image captioning. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (pp. 506–515). Springer.

14. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). CheXclusion: Fairness gaps in deep chest X-ray classifiers. Pacific Symposium on Biocomputing 2021, 232–243.

15. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Nature Communications, 11, 2350.

16. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144.

17. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318.

18. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

19. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.

20. Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., ... & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 1(6), e271–e297.

21. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

22. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).

23. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

24. U.S. Food and Drug Administration. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.

25. European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.