

Mamba-Enhanced Vision-Language Framework for Intelligent Polyp Detection and Clinical Endoscopy Report Generation

Hector Page

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
hectorpage@ucf.edu

Brendan C. Perry

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
helloworld@uc.edu

Abstract

Colorectal cancer prevention depends heavily on the reliable identification and removal of precancerous polyps during endoscopic examination. Although deep learning has improved polyp detection and segmentation, many clinical systems remain narrowly task-oriented and do not address the interpretive documentation workload that follows a procedure. This paper presents a system-level analysis of a Mamba-enhanced vision-language framework that unifies intelligent polyp detection with structured clinical endoscopy report generation. The framework combines selective state space modeling for efficient long-range visual encoding with a vision-language alignment module and a constrained language decoder. The design is discussed from structural, infrastructural, and governance perspectives rather than as a purely algorithmic contribution. The analysis examines how linear-time sequence modeling can reduce memory burdens in long endoscopic video streams while preserving global context for polyp localization. It further considers how language generation can be coupled to visual evidence to produce preliminary reports that support rather than replace clinical judgment. Key trade-offs are addressed across detection sensitivity, false positive burden, language grounding, hallucination control, interoperability with electronic health records, and regulatory compliance. The paper also evaluates fairness risks arising from dataset composition and deployment heterogeneity, and it proposes governance mechanisms that may support safer translation into real clinical environments. The discussion emphasizes that the value of such a framework lies less in isolated benchmark performance than in its ability to operate within an accountable, auditable, and sustainable sociotechnical infrastructure.

Keywords

polyp detection; endoscopy report generation; vision-language models; state space models; Mamba architectures; medical image segmentation; clinical decision support; deployment governance.

1. Introduction

Colorectal cancer remains one of the most preventable malignancies when precancerous lesions are identified and resected during colonoscopy, yet detection rates vary with operator experience, fatigue, bowel preparation quality, and imaging conditions. Machine learning systems have attempted to reduce this variability by providing real-time visual cues, but many existing approaches focus only on bounding boxes or segmentation masks without connecting

visual findings to clinical documentation. Recent advances in selective state space sequence modeling [1] offer an alternative to conventional attention mechanisms by providing linear-time sequence processing while retaining long-range context. In visual domains, adaptations of this line of work [2] have demonstrated that high-resolution feature extraction can be maintained without the quadratic memory cost associated with transformer-based vision models. These developments create a promising foundation for endoscopy systems that must process long video streams and high-resolution mucosal imagery in near real time.

Parallel progress in medical image segmentation has produced specialized state space architectures [3] that can model global context across an entire frame, a property particularly relevant for polyps that vary in size, texture, and boundary definition. At the same time, contrastive vision-language models [4] and medical vision-language assistants [5] have shown that aligning visual and textual representations can support clinically meaningful language outputs. These capabilities motivate an integrated framework that both detects polyps and generates preliminary endoscopy reports. The integration is not straightforward, because the requirements of real-time detection and coherent language generation differ in latency, modularity, and uncertainty representation. A combined system must therefore be designed as an infrastructure that coordinates visual perception, language generation, clinical terminology, and user interaction.

Polyp detection systems based on deep convolutional models have been validated in clinical settings [6], and benchmark datasets [7] have supported reproducible evaluation. Recent detection studies based on updated convolutional architectures [8] and dynamic scanning segmentation networks [9] indicate continued gains from architectural refinement. However, most systems remain task-specific, and their integration with report generation, clinical documentation, and governance structures remains incomplete. This paper addresses that gap at the system level, examining structural trade-offs across the full perception-documentation pipeline. The aim is not to claim that any single model component is superior in isolation, but to analyze how architectural choices, deployment constraints, and policy considerations interact in a practical clinical setting.

Transformer architectures [10] and pretrained language models [11] have established the value of large-scale self-supervised representation learning, but they also introduce infrastructure and latency challenges in hospital environments. The proposed framework therefore combines selective state space models for visual encoding with a vision-language alignment module and a constrained language decoder. The discussion focuses on structural trade-offs, deployment pathways, fairness, and policy implications, reflecting the need to treat clinical artificial intelligence as a socio-technical system rather than an isolated algorithmic artifact.

2. Background and Related Work

Large language models have evolved rapidly from general-purpose text predictors to adaptable instruction-following systems. Open foundation models [12] and scaling studies [13] demonstrate that language generation can be adapted to specialized domains when paired with appropriate instruction tuning [14]. These methods, however, require medical grounding to avoid fluent but inaccurate outputs. In clinical documentation, fluency without visual evidence can produce dangerous statements about lesion location, size, or morphology. Consequently, any report generation module must be tightly coupled to visual representations and constrained by a clinical ontology.

Visual foundation models such as the segment anything architecture [15] illustrate the benefits of promptable visual backbones, yet their use in endoscopy requires task-specific adaptation. Classical segmentation architectures [16] and residual feature extractors [17] remain influential because they provide stable, interpretable baselines that can be more easily validated than very large models. Optimization methods [18] further shape the reproducibility of deployed models by reducing sensitivity to learning rate and weight decay choices. In the endoscopy setting, these stability considerations are especially important because institutions may have different imaging equipment, lighting conditions, and patient populations.

State space models introduce a different computational profile from both convolutional and attention-based architectures. By maintaining a hidden state that is updated sequentially, such models can process high-resolution frames with memory scales linearly in sequence length. This property is attractive for colonoscopy video, where a single procedure may contain tens of thousands of frames. However, linear-time processing is not inherently superior for all clinical tasks. Polyps are often small and localized, requiring strong local feature discrimination as well as global context. The design challenge is to combine the efficiency of state space models with the local precision of convolutional operators and the promptable flexibility of foundation models.

3. System Architecture and Vision-Language Integration

The proposed framework consists of four main subsystems: a Mamba-based visual encoder, a polyp detection and segmentation head, a vision-language projection module, and a clinically constrained language decoder. The visual encoder uses selective state space layers [1] and bidirectional scanning [2] to aggregate spatial context across each frame. Unlike a conventional U-Net [16], the Mamba encoder can maintain global context across high-resolution frames with lower memory, which is valuable during long procedures. The detection head operates on multi-scale features to localize candidate polyps, while the segmentation head refines boundaries for documentation and potential follow-up planning.

A projection layer maps pooled visual representations into a textual embedding space, enabling the decoder to attend to visual features without retraining the full vision encoder. This separation supports modular governance updates. For example, the visual encoder may be updated when new imaging devices are installed without requiring retraining of the language model. Similarly, the language decoder can be improved for terminology coverage without altering detection performance. This modularity is important because clinical software must evolve incrementally under regulatory oversight, and tightly coupled end-to-end models can make change management unnecessarily difficult.

The report generation component is organized around structured sections such as indication, preparation quality, findings, and recommendations. Rather than generating free-form prose from scratch, the decoder is constrained by a template grammar and an ontology of endoscopic findings. The visual features provide evidence for the presence, location, and size of polyps, while the language model converts this evidence into plausible clinical phrasing. The architecture borrows from contrastive vision-language alignment [4] and biomedical instruction tuning [5], but it adds an explicit clinical constraint layer that rejects unsupported statements. The detection and segmentation outputs are therefore not simply inputs to the language model; they serve as evidentiary anchors that the language model must reference when producing each sentence.

4. Mamba-Based Visual Representation and Polyp Detection

The choice of a Mamba-based visual backbone is motivated by the tension between global context and computational efficiency. In traditional convolutional architectures, capturing long-range dependencies requires stacking many layers or using dilated convolutions, which may reduce local precision. Transformer-based vision models provide global context but can become memory-intensive at high spatial resolutions. State space models occupy an intermediate position. By leveraging linear-time sequence processing [1], they can scan flattened image patches while retaining information over long distances. Bidirectional scanning [2] further improves spatial coherence by allowing features to flow in multiple directions across the frame.

In the polyp detection setting, this capability is relevant for distinguishing flat or sessile polyps from surrounding mucosa. Dynamic scanning variants [9] can adapt feature traversal patterns to image content, potentially improving boundary delineation when lesions have irregular shapes. The system can also maintain temporal context by treating consecutive frames as a sequence, allowing the model to associate transient reflections or shadows with stable anatomical structures. This temporal association can reduce false positives caused by artifacts that appear in one frame but not in subsequent frames. However, temporal context can also introduce new failure modes if the model relies excessively on prior frames and becomes less responsive to rapidly emerging lesions.

There are important structural trade-offs in detection sensitivity and false positive control. A system tuned to maximize sensitivity may flag many benign irregularities, increasing cognitive burden and reducing endoscopist trust. A conservative system may miss subtle lesions, particularly those that are flat or partially obscured. The framework addresses this problem by producing confidence-calibrated outputs and by separating localization from report generation. The report module can express uncertainty in a structured way, such as indicating that a region requires further visual inspection. This division of labor allows the detection model to remain alert without forcing the language model to overstate confidence.

The segmentation head also contributes to downstream documentation by producing masks that can be used to estimate size and morphology. Size estimation has clinical importance because larger polyps carry higher malignant potential. A mask-based estimate derived from the image, combined with calibration information from the endoscope, can provide an initial measurement for the report. However, such measurements are approximate and must be verified by the clinician, who can account for distance distortion and camera angle. The system therefore treats segmentation as an assistive measurement rather than a definitive diagnostic output.

5. Clinical Endoscopy Report Generation and Language Alignment

Report generation is a knowledge-intensive task that differs from generic image captioning. A clinically useful endoscopy report must use standardized terminology, follow institutional templates, and distinguish between observed findings and interpreted recommendations. The framework uses pretrained language representations [11] and instruction-following capabilities [14] to generate preliminary text, but it constrains generation through a domain-specific grammar. This hybrid approach attempts to combine the flexibility of large language models with the predictability of structured reporting.

The alignment between visual features and language is achieved through contrastive pretraining [4] and medical vision-language adaptation [5]. During training, paired images and reports teach the model to associate visual patterns with specific descriptors such as

sessile polyp, pedunculated polyp, erythematous mucosa, or ulcerated surface. At inference, the system retrieves relevant visual features and conditions the decoder to produce descriptions that are supported by detected regions. This evidence-driven approach reduces the likelihood of hallucinated findings because each generated sentence must be traceable to an image region or a detected object.

Hallucination remains a central concern. A language model may produce plausible but incorrect statements about lesion location, number, or histology. Histology prediction from white-light imaging is particularly unreliable and should not be included as a direct model output. The framework therefore restricts report statements to visual characteristics and procedural observations, leaving histology to pathology interpretation after resection. This restriction reflects a broader systems principle: the model should document what is observed, not speculate about what cannot be reliably inferred from the current imaging modality.

The language decoder also supports structured data fields that can be mapped to electronic health record systems. In many institutions, free-text reports are less useful for quality measurement and audit than structured fields such as polyp size, location, morphology, and completeness of removal. The framework can generate both a readable narrative and structured fields, reducing the need for manual abstraction. However, this dual output creates a synchronization problem. If the narrative and the structured fields disagree, the clinical user may rely on one and ignore the other. Therefore, the system must maintain a single internal evidence graph from which both outputs are derived.

6. Data Governance, Fairness, and Clinical Policy

Responsible machine learning in health care requires attention to dataset provenance, labeling quality, and deployment monitoring [19]. The clinical translation of artificial intelligence systems demands institutional governance that includes prospective validation, user training, and ongoing audit [20]. A polyp detection and report generation system must be evaluated not only on retrospective test sets but also on its interaction with real clinical workflows. Governance should define who is responsible for reviewing generated reports, how disagreements between model and clinician are recorded, and how errors are escalated.

Legal and ethical challenges [21] complicate the use of artificial intelligence in health care. Liability may arise for missed lesions, false alarms, or incorrect documentation. Developers and health care organizations must clarify whether the system is a decision support tool or a more autonomous actor. In practice, the safest framing is that the system provides suggestions that require clinician confirmation. This framing also affects interface design. The user should be able to see which image regions support a particular report statement, rather than being presented with unverifiable text.

Fairness is a critical but often underexamined issue in medical imaging systems. Models trained primarily on data from one population may perform differently across patient groups due to differences in bowel preparation, anatomical variation, or imaging device characteristics. The framework should include subgroup evaluation across age, sex, ethnicity, and comorbidity profiles where these data are ethically and legally available. Fairness cannot be achieved simply by collecting more data; it requires deliberate attention to the social and institutional conditions under which data are produced and used. Governance committees should include clinicians, data scientists, ethicists, and patient representatives.

Reporting guidelines for artificial intelligence interventions [22] provide a useful starting point for transparency. Such guidelines require clear documentation of model inputs, training

data, evaluation settings, and intended use. For a vision-language framework, additional documentation is needed regarding how language outputs are constrained, how uncertainty is represented, and how the system handles rare or ambiguous findings. These reporting practices can support regulatory review and interinstitutional comparison, but they do not by themselves guarantee safety. Transparency must be accompanied by operational limits on model use and mechanisms for continuous performance tracking.

7. Deployment, Infrastructure, and Operational Sustainability

Deployment of a clinical artificial intelligence system involves more than installing a model on a workstation. Regulatory frameworks for software as a medical device [23] require device classification, quality management, and postmarket surveillance. The visual encoder and the language decoder may fall under different regulatory interpretations depending on whether they provide diagnostic suggestions or merely assist with documentation. This distinction matters because it affects the evidence required for approval and the obligations of the manufacturer after deployment.

The computational infrastructure must accommodate endoscopy suites with variable network connectivity, data storage limitations, and cybersecurity requirements. A cloud-only architecture may simplify model updates but raises privacy concerns and latency risks. An edge deployment may improve responsiveness but complicates monitoring and version control. The Mamba-based design is potentially advantageous for edge deployment because its memory footprint can be smaller than transformer-based alternatives [1], but real-world latency depends on hardware accelerators, input resolution, and the complexity of the decoding process. Operational benchmarking must therefore occur under realistic conditions rather than controlled laboratory settings.

Sustainability also includes technical maintenance and institutional capacity. Models can drift as imaging devices change or as patient populations shift. Continuous monitoring requires access to representative validation data, which may conflict with privacy protection and data minimization principles. Federated evaluation frameworks can provide some oversight without centralizing raw data, but they introduce additional governance complexity. Institutions also need personnel who understand both the technical system and the clinical workflow. Without this human infrastructure, even a well-designed model may become an unmanaged risk.

The environmental footprint of artificial intelligence is increasingly relevant to sustainable health care. Large-scale training runs consume significant energy and computational resources. Although the proposed framework is designed for efficiency at inference, its training may still rely on large pretrained vision and language components. A full lifecycle assessment should consider not only model accuracy but also energy consumption, hardware replacement cycles, and data storage. Efficiency-aware architecture choices such as selective state space models [2] may help reduce operational energy demands, but they do not eliminate the need for institutional sustainability planning.

8. Evaluation Strategy and Robustness Considerations

Evaluation of a combined detection and report generation system requires multiple levels of assessment. Technical benchmarks can measure detection sensitivity, segmentation overlap, and language quality, but these metrics do not capture clinical usefulness. The system must be evaluated in simulated workflows with endoscopists under realistic time pressure and varying

case complexity. User experience studies should examine whether generated reports reduce documentation time, improve completeness, or introduce new forms of cognitive bias.

Robustness to distribution shift is particularly important in gastrointestinal imaging. Bowel preparation quality varies widely, and poor preparation can obscure polyps or create false positives. The framework should be assessed on stratified data representing different preparation scales, anatomical locations, and polyp morphologies. Stress testing should include frames with specular reflections, motion blur, surgical clips, and retained fluid. A robust system should be able to indicate when its confidence is low and request clarification from the human operator, rather than forcing a deterministic output under uncertain conditions.

The interaction between visual detection and report generation also creates failure modes that do not exist in single-task systems. A correct detection with an incorrect textual interpretation can be more dangerous than a missed detection because it may be copied into the medical record without verification. The system should include an audit trail that links each report statement to the evidence from which it was derived. This traceability enables post hoc review and provides a basis for clinician trust. It also complicates the user interface, which must present enough evidence without overwhelming the clinician during a procedure.

Prospective clinical evaluation remains essential. Retrospective performance on curated datasets may overstate real-world utility because it avoids the selection biases and workflow interruptions present in practice. Clinical studies should measure not only polyp detection rates but also documentation accuracy, report review time, and unanticipated harms. The regulatory and governance frameworks discussed earlier [19], [22], [23] provide guidance for conducting such studies in an ethically responsible manner. The overall objective should be to demonstrate that the system improves the quality of care without increasing clinician burden or masking diagnostic uncertainty.

9. Conclusion

This paper has examined a Mamba-enhanced vision-language framework for intelligent polyp detection and clinical endoscopy report generation from a systems perspective. The framework combines selective state space modeling for efficient visual representation, segmentation-based evidence extraction, and constrained language generation. Rather than treating polyp detection and report writing as separate technical tasks, the design integrates them through an evidence graph that supports both structured data and narrative text. This integration offers potential benefits for documentation efficiency and clinical audit, but it also introduces new risks related to hallucination, fairness, liability, and infrastructure maintenance.

The analysis suggests that architectural efficiency is a necessary but insufficient condition for clinical success. A model may achieve strong benchmark performance yet fail in deployment because of incompatible workflows, unclear accountability, or unmanaged distribution shift. The Mamba backbone offers practical advantages in memory and long-range context, and detection studies based on updated convolutional methods [8] and dynamic segmentation networks [9] illustrate the continuing importance of architectural refinement. However, the broader contribution of the framework lies in its attention to governance, modularity, and human oversight.

Future work should prioritize prospective studies, standardized reporting, and interinstitutional collaboration. Evaluation should extend beyond lesion detection to include

the accuracy and traceability of generated reports. As clinical artificial intelligence becomes more integrated into gastroenterology, the field must develop governance structures that respect patient privacy, promote fairness, and ensure that automated systems remain accountable to the clinicians and patients they are intended to serve.

References

1. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
2. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417.
3. Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.
4. 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). PMLR.
5. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., ... & Gao, J. (2024). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36.
6. Wang, P., Xiao, X., Glissen Brown, J. R., Berzin, T. M., Tu, M., Xiong, F., ... & Liu, X. (2018). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature Biomedical Engineering, 2(10), 741-748.
7. Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: A segmented polyp dataset. In Proceedings of the 26th International Conference on Multimedia Modeling (pp. 451-462). Springer.
8. Tang, Y., Chang, C., Zhao, Z., Gu, Z., & Li, Z. (2026, March). Intelligent Detection and Segmentation of Gastrointestinal Polyps Based on YOLOv11 Deep Learning. In 2026 9th International Conference on Advanced Algorithms and Control Engineering (ICAACE) (pp. 825-828). IEEE.
9. Zhang, Z., Ma, Q., Zhang, T., Chen, J., Zheng, H., & Gao, W. (2025). Switch-UMamba: Dynamic scanning vision Mamba UNet for medical image segmentation. Medical Image Analysis, 103792.
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
11. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).
12. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

13. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (pp. 1877-1901).
14. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015-4026).
16. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer.
17. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).
18. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.
19. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337-1340.
20. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
21. Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare* (pp. 295-336). Academic Press.
22. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26(9), 1364-1374.
23. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. U.S. Department of Health and Human Services.
24. 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.
25. Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., ... & Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5), 820-838.