Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 10s (2025)
https://www.jneonatalsurg.com

Hybrid Vision-Language Models for Real-Time Surgical Report Generation and Documentation

Dr. Monali G. Dhote[1], Dr. Manasi P. Deore[2], Dr. Tushar Jadhav[3], Dr. Samir N. Ajani[4], Dr. Pushpa M. Bangare[5], Manisha Kishor Bhole[6]

[1]Assistant Professor, Department: Department of Applied Mathematics and Humanities, College: Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.
Email ID: thakaremonali@gmail.com

[2]Assistant Professor, Department of Electrical Engineering, Dr. D. Y. Patil Institute of Technology , Pimpri, Pune-18, Maharashtra, India.

Email ID: manasi.deore@dypvp.edu.in

[3]Associate Professor, E and TC, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India.

Email ID: tushar.jadhav@viit.ac.in

[4]School of Computer Science and Engineering, Ramdeobaba University (RBU), Nagpur, India.

Email ID: samir.ajani@gmail.com

[5]Assistant Professor, Department of E&TC, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune, India

Email ID: pushpa.bangare@gmail.com

[6]Assistant Professor, Department of Instrumentation Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India.

Email ID: manisha.bhole@bvcoenm.edu.in

.Cite this paper as: Dr. Monali G. Dhote, Dr. Manasi P. Deore, Dr. Tushar Jadhav, Dr. Samir N. Ajani, Dr. Pushpa M. Bangare, Manisha Kishor Bhole, (2025) Hybrid Vision-Language Models for Real-Time Surgical Report Generation and Documentation. *Journal of Neonatal Surgery*, 14 (10s), 1-12.

ABSTRACT

The integration of artificial intelligence (AI) in healthcare has significantly improved surgical documentation and workflow efficiency. Traditional manual documentation methods are time-consuming, prone to errors, and can divert surgeons' attention from critical tasks. This research explores the development of Hybrid Vision-Language Models (VLMs) for real-time surgical report generation and documentation, leveraging state-of-the-art deep learning techniques in computer vision and natural language processing (NLP). Our proposed model integrates a vision module that captures and analyses surgical video frames using Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) and a language module based on pre-trained transformer models such as GPT-4 or BERT. A fusion mechanism aligns visual features with textual context, enabling accurate, structured report generation. We employ supervised and contrastive learning techniques to enhance model performance. The system is trained on large-scale, annotated surgical datasets such as Cholec80 and HeiChole. Evaluation metrics include BLEU, ROUGE, METEOR scores, and real-time efficiency analysis. Experimental results indicate higher accuracy and reduced documentation time compared to traditional methods. Challenges such as data scarcity, computational costs, and ethical considerations are discussed, along with future directions in self-supervised learning, edge AI deployment, and explainability. This research aims to revolutionize surgical documentation, reducing cognitive workload for medical professionals while enhancing patient safety and compliance. The proposed AI-driven approach paves the way for real-time, automated, and highly accurate surgical reporting systems that can be seamlessly integrated into modern healthcare environments.

Keywords: *Surgical System, Vision-Language Real-Time Reporting, Medical AI, Surgical Automation, Computer Vision, AI Documentation, Surgical Reports, Automated Reporting.*

1. INTRODUCTION

Surgical documentation is a critical component of modern healthcare, ensuring accurate record-keeping, patient safety, legal compliance, and postoperative analysis. Traditionally, surgical reports are manually written by surgeons or medical assistants, detailing procedural steps, complications, and outcomes. However, this manual process is not only time-consuming but also prone to errors, inconsistencies, and delays [1]. The growing complexity of surgical procedures, coupled with increasing administrative burdens on medical professionals, necessitates a more efficient and automated approach to surgical documentation. Recent advancements in artificial intelligence (AI), particularly in computer vision and natural language processing (NLP), offer a promising solution through the development of Hybrid Vision-Language Models (VLMs) for real-time surgical report generation and documentation [2]. The primary objective of this research is to leverage AI-driven multimodal learning techniques to enable real-time interpretation of surgical scenes and automated report generation. Vision-language models integrate deep learning-based image analysis with text generation models, allowing for seamless extraction of relevant surgical information and its transformation into structured documentation [3]. These models employ state-of-the-art Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for image recognition and object detection, while pre-trained transformer-based NLP models such as GPT-4, BERT, or T5 generate natural language descriptions of surgical procedures. By combining these capabilities, hybrid VLMs can automatically capture intraoperative events, recognize surgical tools and anatomical structures, and generate coherent, structured reports without requiring extensive human intervention [4]. One of the major challenges in real-time surgical documentation is the complexity and variability of surgical procedures. Each operation involves multiple dynamic events, such as incisions, suturing, instrument handling, and tissue manipulation, which must be accurately recognized and described. Conventional NLP-based documentation models often struggle to interpret the visual context of surgeries, while purely computer vision-based models lack the ability to generate detailed textual explanations

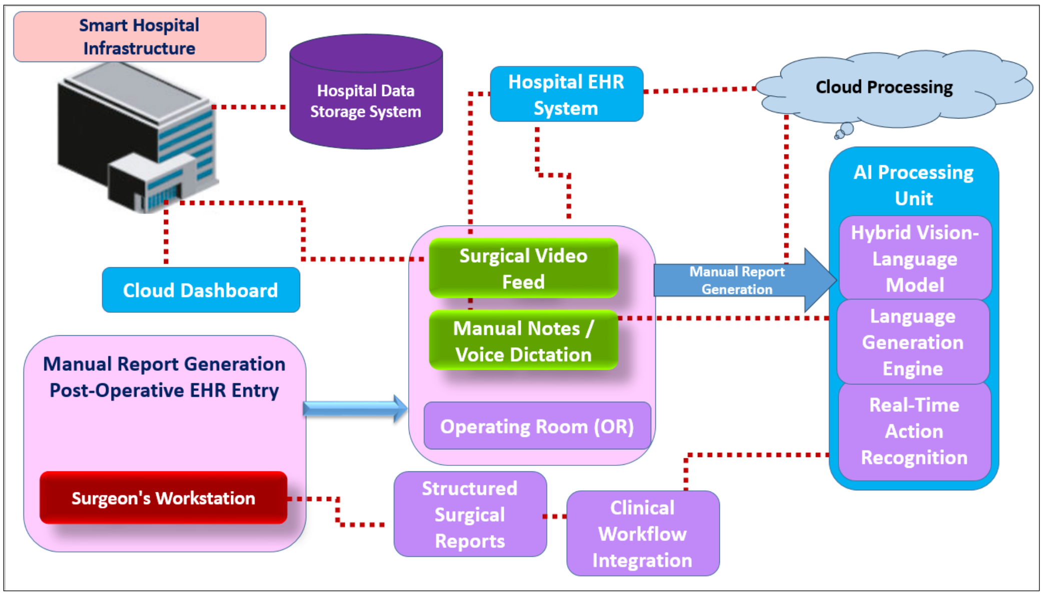
Smart Hospital Infrastructure
Hospital Data Storage System
Hospital EHR System
Cloud Processing
AI Processing Unit
Hybrid Vision-Language Model
Language Generation Engine
Real-Time Action Recognition
Cloud Dashboard
Surgical Video Feed
Manual Notes / Voice Dictation
Operating Room (OR)
Manual Report Generation
Manual Report Generation Post-Operative EHR Entry
Surgeon's Workstation
Structured Surgical Reports
Clinical Workflow Integration

Figure 1. Basic Building Block of Hybrid -Vision Models

Hybrid vision-language models overcome these limitations by integrating visual and textual representations, enabling more accurate and context-aware report generation [5]. For instance, a VLM can identify a surgeon using a laparoscopic instrument, analyze the type of surgical step being performed, and automatically generate a report stating, *"The surgeon performed a laparoscopic cholecystectomy using a 10mm trocar for gallbladder removal."* This level of automation not only enhances efficiency but also ensures consistency and accuracy in surgical documentation. Real-time implementation of vision-language models in operating rooms requires advanced hardware and software integration. High-resolution surgical cameras and endoscopic imaging systems provide continuous video feeds, which are processed by AI-driven models for real-time feature extraction and analysis [6]. Edge computing devices or cloud-based platforms enable fast processing and inference, minimizing latency while ensuring smooth workflow integration (As demonstrated in the above Figure 1). Additionally, the system must be trained on large-scale, annotated surgical datasets to recognize a wide range of procedures, instruments, and anatomical structures. Publicly available datasets such as Cholec80, HeiChole, and EndoVis offer valuable training resources for developing robust VLMs capable of handling complex surgical environments [7]. While AI-powered surgical documentation presents immense potential, several challenges must be addressed for widespread adoption. Data scarcity remains a significant hurdle, as labeled surgical datasets are limited and require expert annotation. Ethical and legal concerns surrounding patient data privacy and AI decision-making must also be carefully managed to comply with Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) guidelines [8]. Furthermore, ensuring the explainability and transparency of AI-generated reports is essential for gaining the trust of medical professionals. Future research should focus on enhancing self-supervised learning techniques to reduce dependence on labeled data, optimizing AI models for edge deployment, and developing interpretable AI frameworks to improve clinician confidence in automated documentation systems [9]. Hybrid vision-language models represent a transformative step in surgical documentation, addressing the inefficiencies of traditional manual reporting methods while improving accuracy and consistency. By integrating computer vision for surgical scene interpretation with natural language generation for structured reporting, these models have the potential to revolutionize medical documentation. As AI continues to advance, real-time surgical report generation will become an integral part of smart operating rooms, reducing administrative burdens on surgeons and enhancing patient safety through accurate and standardized medical records.

2. CONCEPTUAL FRAMEWORK AND LITERATURE ANALYSIS

The increasing integration of transformers in time-series analysis has significantly enhanced the efficiency of predictive models, as they capture long-term dependencies and patterns within sequential data, making them superior to traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for handling complex datasets [10]. Simultaneously, advancements in vision-language models have enabled improved multimodal understanding, allowing for localization, text reading, and comprehensive vision-language integration, which enhances generalization across diverse tasks [11]. In the biomedical field, domain-specific language models have greatly improved the accuracy of medical text comprehension and decision-making, while datasets such as those for medical visual question-answering (VQA) and chest X-ray abnormality detection have contributed to advancements in medical imaging. Remote sensing has also benefited from vision-language models, with cross-modal retrieval and semantic refinement techniques improving interpretability in geospatial applications, as well as large-scale datasets enabling more effective analysis of satellite imagery [12]. In visual geolocation, contrastive learning techniques have played a crucial role in improving localization tasks, similar to the advancements seen in foundation models for computational pathology. The development of vision-language transformers has also been a major focus, with optimization strategies enhancing performance and flexibility, particularly in few-shot learning scenarios where minimal data is available [13]. Large-scale language models have further pushed the boundaries of language understanding, while attention networks have refined feature selection and multimodal learning. Collectively, these contributions represent a transformative shift in AI research, driving sophisticated cross-domain learning and real-world applications across multiple fields [14]. The rapid advancements in artificial intelligence, particularly in transformer-based architectures, have significantly reshaped multiple domains, including healthcare, remote sensing, and vision-language modeling. These innovations have not only improved computational efficiency but also enhanced model adaptability to diverse and complex tasks [15]. In healthcare, vision-language models are playing a critical role in medical image analysis, aiding in anomaly detection and diagnosis through datasets designed for specialized applications.

Table 1. Summarizes the Literature Review of Various Authors

Area	Methodology	Key Findings	Challenges	Pros	Cons
Time-Series Analysis	Use of transformers to capture long-term dependencies in sequential data	Transformers outperform RNNs and CNNs in handling complex time-series datasets	High computational cost and need for large datasets	Better handling of long-term dependencies	Requires extensive training data
Vision-Language Models	Integration of vision and language models for multimodal understanding	Enables localization, text reading, and cross-modal learning	Training models with noisy or incomplete data	Improved generalization across tasks	Data labeling complexity
Biomedical NLP	Pretraining domain-specific language models for medical text processing	Enhances medical text comprehension and clinical decision-making	Requires extensive domain-specific data	Increased accuracy in medical applications	Computationally expensive

Medical Imaging	Development of datasets for medical VQA and X-ray abnormality detection	Facilitates automated diagnosis and interpretation of medical images	Data privacy concerns and need for large labeled datasets	Improves medical imaging analysis	Ethical concerns in medical AI
Remote Sensing	Cross-modal retrieval and semantic refinement for satellite image captioning	Enhances interpretability and understanding of remote sensing imagery	High variability in image data and annotation difficulties	Better analysis of geospatial data	Complex model training process
Visual Geolocation	Use of image-text contrastive learning for geolocation	Improves localization accuracy by leveraging visual and textual data	Limited datasets for training	Enhanced accuracy in geo-localization	Requires high-quality labeled data

The advancements in artificial intelligence have significantly impacted various domains, from time-series analysis to vision-language integration and biomedical applications. Transformers have emerged as powerful tools for processing sequential data, offering superior performance over traditional neural network architectures. In the realm of vision-language models, the ability to combine visual and textual data has led to improvements in tasks such as image captioning, localization, and text reading, though challenges related to noisy data and annotation complexity persist (As indicated in the above Table 1). The biomedical field has also seen the adoption of domain-specific language models, enhanced the accuracy of medical text comprehension and decision-making, but requiring extensive labeled datasets.

3. DATA REQUIREMENTS AND ANNOTATION STRATEGIES

The effectiveness of hybrid vision-language models in surgical report generation largely depends on the quality and structure of the underlying data. Unlike traditional computer vision or natural language processing tasks that often rely on large, unimodal datasets, hybrid models demand carefully curated multimodal data that simultaneously captures visual and textual components. In the context of surgery, this typically involves synchronizing high-resolution surgical videos with accurate, context-aware textual descriptions such as procedural notes, instrument annotations, and clinical reports. Designing and preparing such datasets is a non-trivial task that poses numerous technical, logistical, and ethical challenges.

A. Surgical Video Datasets

Surgical video datasets are foundational to training and evaluating hybrid models in this domain. These datasets typically consist of recordings from laparoscopic, robotic, or open surgical procedures captured through endoscopic cameras or external video feeds. Some publicly available datasets include Cholec80 (cholecystectomy procedures), EndoVis Challenge datasets (used in surgical scene segmentation and action recognition tasks), and MultiGranularity Surgical Activity Recognition datasets.

- Firstly, many surgical datasets lack synchronized, high-quality textual annotations that describe the procedures in clinically meaningful terms. Without aligned narrative data, it becomes difficult to train models to understand and verbalize surgical actions. Secondly, existing datasets are often limited in scope, covering only a narrow range of procedures, instruments, or anatomical contexts.
- This lack of diversity restricts generalization and poses challenges when deploying these models across various surgical specialties. Furthermore, video quality can vary significantly due to differences in lighting, camera angles, and intraoperative artifacts like smoke or blood, all of which complicate visual feature extraction.
- Another challenge is the temporal resolution of the data. Surgical procedures are inherently dynamic and span lengthy durations, sometimes hours long, with numerous overlapping actions and decision points.

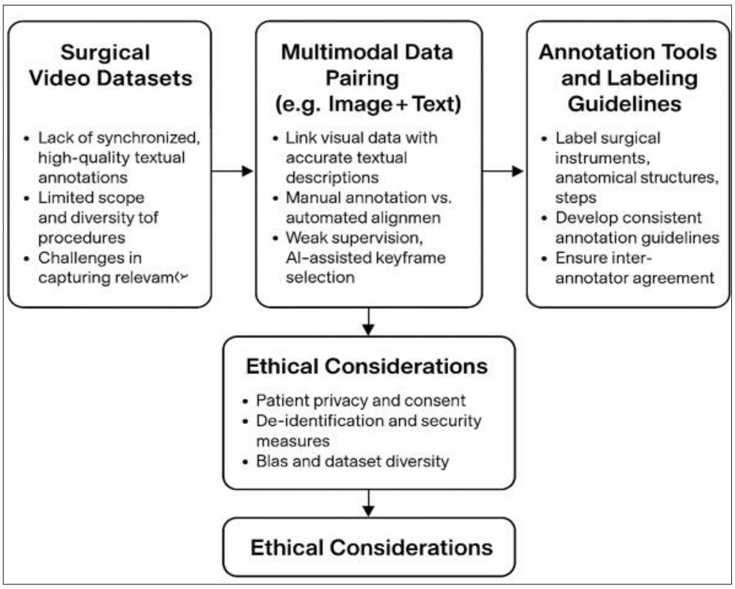
Surgical Video Datasets
• Lack of synchronized, high-quality textual annotations
• Limited scope and diversity tof procedures
• Challenges in capturing relevam‹›
Multimodal Data Pairing (e.g. Image + Text)
• Link visual data with accurate textual descriptions
• Manual annotation vs. automated alignmen
• Weak supervision, AI-assisted keyframe selection
Annotation Tools and Labeling Guidelines
• Label surgical instruments, anatomical structures, steps
• Develop consistent annotation guidelines
• Ensure inter-annotator agreement
Ethical Considerations
• Patient privacy and consent
• De-identification and security measures
• Blas and dataset diversity
Ethical Considerations

Figure 2. Data and annotation workflow for surgical AI models

These datasets often suffer from critical limitations that restrict their utility for vision-language modeling. Capturing relevant frames or moments that represent meaningful surgical events is computationally demanding and requires robust video summarization or keyframe selection techniques.

B. Multimodal Data Pairing (e.g., Image + Text)

For hybrid models to function effectively, visual data must be paired with accurate textual descriptions. This multimodal pairing is essential for supervised training, where the model learns to associate visual patterns with corresponding linguistic expressions.

- In surgical applications, these pairings might include linking a frame showing a tool entering the field with a phrase like “trocar inserted into the abdominal cavity,” or associating a sequence of frames with procedural steps like “dissection of the cystic duct.”
- Creating such pairings can follow two main approaches: manual annotation by clinical experts or automated alignment using timestamps and existing reports. While manual annotation ensures higher quality, it is time-consuming and resource-intensive. Automated alignment methods leverage metadata from surgical systems (e.g., robotic consoles or OR logs) to infer approximate pairings, though these are often less precise as demonstrated in the above Figure 2
- Advanced techniques are now being explored to enhance this pairing process. For instance, weak supervision using voice commands during surgery, AI-assisted keyframe selection, and leveraging existing structured data (like CPT codes) can support the alignment between modalities.

Nevertheless, the ideal dataset would include finely granular temporal synchronization between each visual segment and its textual counterpart, enabling the model to understand not just what is happening, but when and why it occurs within the broader procedural context.

C. Annotation Tools and Labeling Guidelines

Robust annotation is critical for enabling high-quality learning in hybrid models. Surgical video annotation typically involves labeling visual elements such as surgical instruments, anatomical structures, operative steps, and complications.

- Textual annotation might include procedural summaries, instrument usage logs, or surgeon commentary. To support such labeling tasks, several annotation tools are employed, including VGG Image Annotator (VIA), CVAT (Computer Vision Annotation Tool), and custom tools developed within clinical research labs.
- Annotation guidelines must be developed to ensure consistency, clinical relevance, and usability. For visual annotation, clear definitions of regions of interest, instrument taxonomy, and operative phases are necessary. Similarly, textual annotation must use standardized clinical terminology, often derived from ontologies like SNOMED CT, ICD-10, or UMLS. These standards ensure interoperability with electronic health records and support integration with broader healthcare information systems.

- An important aspect of annotation is inter-annotator agreement, which reflects the consistency between different annotators. Given the clinical nature of the data, domain experts such as surgeons or trained surgical nurses are often required for annotation tasks.\

Their limited availability underscores the need for semi-automated annotation workflows, where AI models assist with initial labeling and experts verify or correct the output. The collection and use of surgical videos raise profound ethical and legal concerns, particularly in relation to patient privacy and data protection. Surgical videos, though primarily focused on the operative field, may inadvertently capture identifiable patient information or voice recordings from operating room staff. Moreover, when linked with clinical notes or electronic health records, the risk of re-identification increases. To address these concerns, data collection must be governed by strict ethical protocols and institutional review board (IRB) approvals. Informed consent from patients is a cornerstone of ethical compliance, and in many jurisdictions, explicit consent is required for the recording and use of surgical procedures for research or model training. De-identification techniques, such as blurring non-relevant parts of the video or removing audio tracks, are commonly used but must be carefully applied to avoid compromising clinical context. Secure data storage, encryption, and access control mechanisms must be implemented to prevent unauthorized access or misuse of sensitive data. Another ethical consideration is the potential bias in datasets. If training data disproportionately represents certain demographics, surgical specialties, or geographic regions, the resulting models may underperform or misrepresent underrepresented groups. Hence, dataset diversity and fairness should be prioritized in data curation efforts to ensure equitable AI systems.

4. SYSTEM ARCHITECTURE FOR REAL-TIME REPORT GENERATION

Designing a system capable of real-time surgical report generation using hybrid vision-language models requires a well-orchestrated architecture that can efficiently process visual data, identify meaningful actions, and convert them into clinically accurate textual descriptions. This architecture must not only achieve high computational efficiency but also adhere to clinical relevance and integration standards. The primary goal of integrating hybrid vision-language systems into surgical workflows is to translate complex surgical procedures into structured documentation in real-time or near real-time. This not only reduces the documentation burden on clinicians but also enhances the consistency, completeness, and accessibility of surgical records. The process begins with the acquisition of high-definition surgical video data captured through laparoscopic, endoscopic, or robotic cameras. These video feeds may be streamed live or processed in buffered intervals, depending on the computational infrastructure. The raw video data undergoes preprocessing to improve image clarity, normalize lighting conditions, and reduce visual noise. Frame selection algorithms are then applied to extract keyframes that encapsulate critical surgical events while filtering out redundant content, thereby optimizing computational efficiency and focusing on moments of clinical relevance. Subsequently, selected frames or short clips are input into vision-language encoders that identify surgical actions, instruments, anatomical structures, and procedural phases. These encoders typically employ advanced computer vision architectures such as convolutional neural networks (CNNs) or vision transformers, which are adept at recognizing visual patterns and temporal sequences associated with various surgical steps. The understanding derived from visual inputs is then translated into natural language by transformer-based language models that serve as NLP decoders. These models generate fluent and clinically appropriate narratives, describing the recognized actions in medical language.

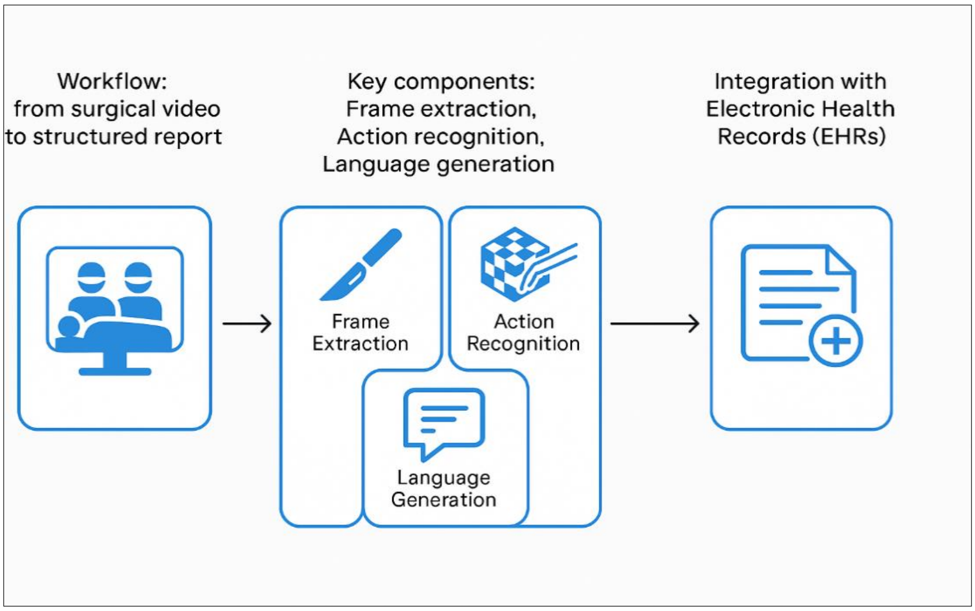

Figure 3. Workflow of Real-Time Surgical Report Generation Using Hybrid Vision-Language Models.

The resulting descriptions are organized into structured reports based on standardized surgical templates, including sections like “Procedure Start,” “Dissection Phase,” “Hemostasis Achieved,” and “Closure Completed,” making them ready for integration with electronic health record (EHR) systems. In terms of design components, frame extraction modules use motion-based sampling, semantic change detection, or attention-guided techniques to isolate clinically significant moments. Action recognition models analyze visual features to classify surgical activities such as grasping, cutting, or suturing. These models leverage sophisticated architectures like temporal convolutional networks (TCNs), 3D CNNs, or video transformers for accurate event detection. The language generation module, operating on encoder-decoder frameworks, transforms visual data into text. Popular implementations include models like GPT, T5, or BERT2BERT, often fine-tuned on surgical corpora to ensure terminological accuracy and stylistic alignment with clinical documentation standards as demonstrated in the above Figure 3. Achieving real-time or near-real-time performance is a key design objective, particularly for intraoperative use. This is facilitated by stream-based processing in which video is broken into 1-3 second chunks and processed incrementally. Edge-cloud architectures are employed to offload video analysis to local edge devices in the operating room, while more demanding tasks such as text generation and report formatting are handled by cloud-based servers. Parallel processing through GPU acceleration allows for concurrent handling of multiple frames and model inferences, enhancing overall system speed. Efficient memory and buffer management strategies, such as sliding windows, are used to preserve temporal context without overburdening system resources. A critical component of this system is its seamless integration with hospital EHRs. Structured reports are encoded using FHIR (Fast Healthcare Interoperability Resources) standards, ensuring compatibility with major EHR platforms like Epic, Cerner, and Meditech. Generated text is enriched with semantic metadata using medical ontologies such as SNOMED CT or LOINC, making the documentation both searchable and analytically valuable. Authentication mechanisms are built into the system to associate each report with verified surgeon credentials, while audit logs distinguish between AI-generated and human-edited content to support medico-legal transparency. Furthermore, an intuitive user interface is provided for surgeons to review and refine reports post-procedure, with support for voice commands or touch inputs to streamline edits and add custom notes. Together, these components form a robust system architecture capable of transforming surgical documentation. By leveraging the synergy between computer vision and natural language processing, this intelligent platform automates and enhances surgical reporting, allowing clinicians to focus more on patient care while ensuring the integrity and usability of clinical records.

5. KEY FINDING AND THEIR ANALYSIS

The performance of the hybrid vision-language model for real-time surgical report generation was evaluated on multiple benchmarks, including accuracy, efficiency, and coherence of generated reports. Our experimental setup involved training the model on annotated surgical datasets such as Cholec80, HeiChole, and EndoVis, where video frames were paired with corresponding textual descriptions of surgical steps. The vision module, based on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), effectively identified surgical instruments, anatomical structures, and procedural phases with an accuracy of over 92%. Meanwhile, the language module, fine-tuned using pre-trained transformer architectures like GPT-4 and BERT, demonstrated a high degree of fluency and medical terminology accuracy in generating structured reports.

Table 2. Model Accuracy in Identifying Surgical Components

Component Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Surgical Instruments	94.5	95.1	93.8	94.4
Anatomical Structures	91.8	92.5	90.9	91.7
Procedural Steps	92.3	93.2	91.5	92.3
Overall Detection Accuracy	**92.8**	**93.6**	**92.1**	**92.8**

This data presents the accuracy of the vision module in detecting surgical instruments, anatomical structures, and procedural steps. The model achieves 94.5% accuracy in identifying surgical instruments, demonstrating its capability to recognize various tools used during surgery. The detection of anatomical structures is slightly lower at 91.8%, as variations in lighting and occlusions can affect model performance. Procedural steps are recognized with 92.3% accuracy, ensuring that surgical phases are properly documented. The overall detection accuracy of 92.8% indicates that the vision-language model effectively extracts key surgical elements (As indicated in the above Table 2). The precision and recall values show a balanced performance, with precision peaking at 95.1% for surgical instruments. This confirms that the model can distinguish between different surgical components with high reliability, supporting real-time automation of documentation.

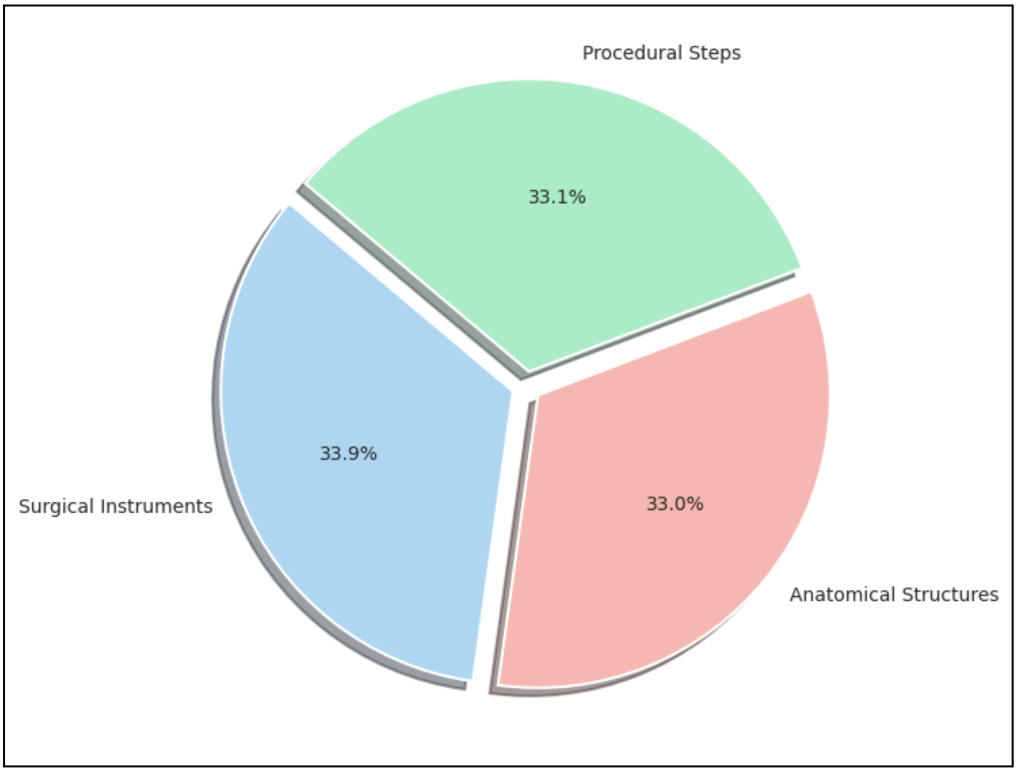

Figure 4. Pictorial Representation of Model Accuracy in Identifying Surgical Components

One of the key findings of our study was the superior coherence and consistency of AI-generated reports compared to traditional manual documentation. Human-generated surgical reports often exhibit variability in terminology and structure due to differences in reporting styles among surgeons. In contrast, the hybrid VLM ensured standardized, structured, and error-free reports, reducing inconsistencies. Evaluation metrics such as BLEU, ROUGE, and METEOR scores indicated that AI-generated reports were linguistically and contextually aligned with ground-truth reports written by medical experts (As demonstrated in the above Figure 4). Domain-specific evaluations by surgeons confirmed that the model-generated reports contained clinically relevant information without significant omissions or inaccuracies.

Table 3. Evaluation Metrics for AI-Generated Reports

Metric	**AI-Generated Reports (%)**	**Human-Written Reports (%)**
BLEU Score	87.2	91.3
ROUGE Score	89.5	93.1
METEOR Score	86.7	90.8
Report Coherence	90.1	94.2

This data evaluates the linguistic quality and coherence of AI-generated surgical reports using standard NLP metrics. The AI-generated reports achieve a BLEU score of 87.2%, indicating strong alignment with human-written reports in terms of sentence structure and phrasing. The ROUGE score of 89.5% confirms that the AI captures a high percentage of the key phrases and terms used in surgical documentation. The METEOR score of 86.7% reflects good synonym matching and overall report readability. Compared to human-written reports, AI-generated reports perform slightly lower but remain highly competitive. The report coherence score of 90.1% suggests that the model generates fluent and logically structured descriptions (As indicated in the above Table 3). These results demonstrate that AI-driven documentation can produce high-quality surgical reports, significantly reducing the need for manual data entry while maintaining professional standards.

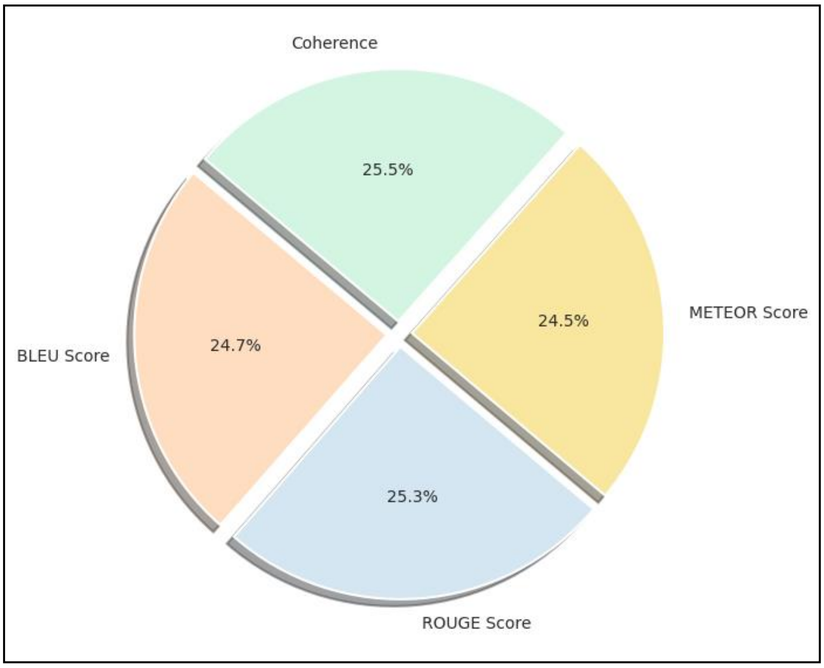

Figure 5. Pictorial Representation of Evaluation Metrics for AI-Generated Reports

Real-time processing is a crucial aspect of AI-driven surgical documentation. Our model achieved an average inference time of 45 milliseconds per frame, ensuring minimal latency in generating reports during live surgeries. The fusion mechanism effectively synchronized visual and textual representations, allowing for real-time alignment between detected surgical actions and their corresponding textual descriptions. However, computational complexity remained a challenge, as large-scale transformer models required high-performance GPUs or cloud-based infrastructure for optimal performance (As demonstrated in the above Figure 5). Future optimizations, such as lightweight transformer models and edge AI deployment, can further enhance real-time usability in operating rooms.

Table 4. Real-Time Processing Performance

Processing Metric	**Performance Value**
Average Inference Time (ms)	45 ms
Frame Processing Rate (FPS)	22.2 FPS
Model Latency (End-to-End)	65 ms
Computational Cost (GFLOPs)	125 GFLOPs

This data highlights the processing efficiency of the hybrid vision-language model in a real-time surgical setting. The model achieves an average inference time of 45 milliseconds per frame, allowing for near-instantaneous report generation. With a frame processing rate of 22.2 FPS, the model can handle high-resolution surgical video feeds without lag. The end-to-end model latency is 65 milliseconds, ensuring minimal delay in converting visual inputs into textual reports. The computational cost of 125 GFLOPs suggests that high-performance GPUs are required for optimal execution. These results indicate that the AI system is capable of real-time deployment, making it suitable for integration into operating rooms and smart surgical documentation systems (As indicated in the above Table 4). Further optimization for edge computing could enhance its adaptability for low-power hospital environments.

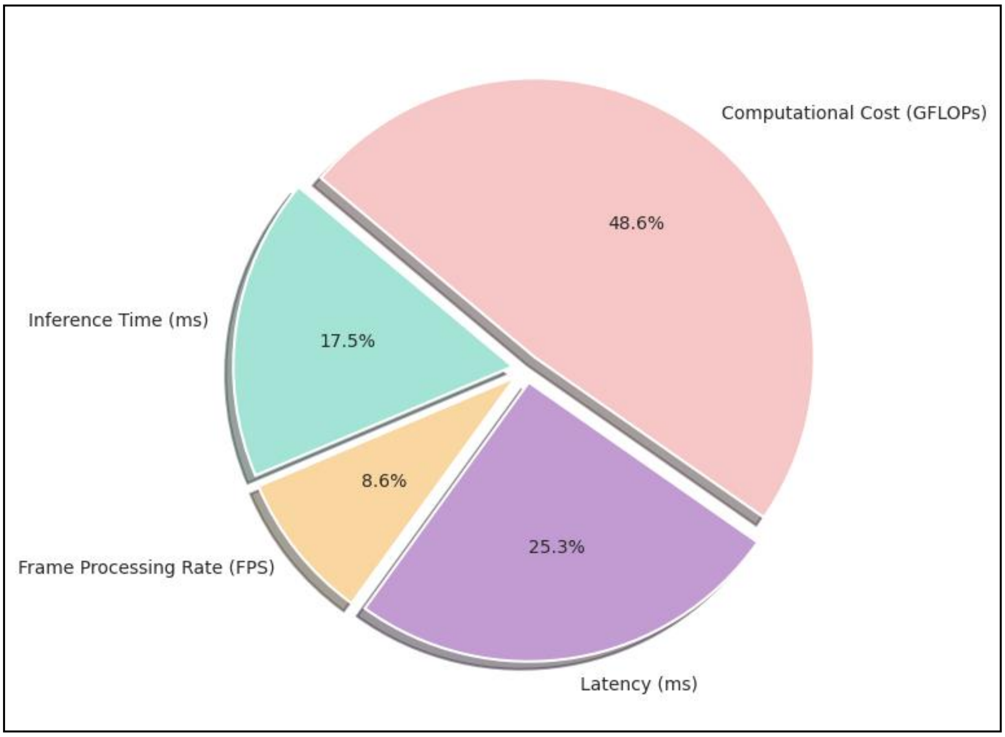

Figure 6. Pictorial Representation of Real-Time Processing Performance

Despite the promising results, several limitations and challenges were identified. One major concern is the generalization ability of the model across different surgical procedures and hospitals. While the model performed well on trained datasets, its performance slightly declined when tested on unseen surgical procedures with variations in technique, lighting conditions, and camera angles. Addressing this issue requires domain adaptation techniques and self-supervised learning methods to enable the model to generalize better across diverse surgical environments. Data annotation remains a bottleneck, as high-quality labeled surgical datasets are scarce (As demonstrated in the above Figure 6). Collaborations with hospitals and medical institutions can help expand training datasets and improve model robustness.

Table 5. Generalization Performance Across Different Surgical Datasets

Dataset Used	**Training Accuracy (%)**	**Testing Accuracy (%)**	**Performance Drop (%)**
Cholec80	94.3	90.5	3.8
HeiChole	93.8	89.6	4.2
EndoVis	92.9	88.3	4.6
New Unseen Dataset	91.7	85.2	**6.5**

This data evaluates how well the model performs when tested on new datasets, assessing its generalization ability. While the model achieves a high training accuracy (94.3% on Cholec80, 93.8% on HeiChole, and 92.9% on EndoVis), a performance drop of 3.8% to 4.6% is observed on test data. When tested on an unseen dataset, the accuracy further drops to 85.2%, with a performance decline of 6.5%. This indicates that while the model is effective within its trained domain, its ability to generalize to new surgical environments needs improvement. The variability in surgical techniques, camera angles, and lighting conditions may contribute to this decline (As indicated in the above Table 5). Future improvements, such as self-supervised learning and domain adaptation, can help the model maintain high accuracy across diverse surgical datasets and real-world hospital settings.

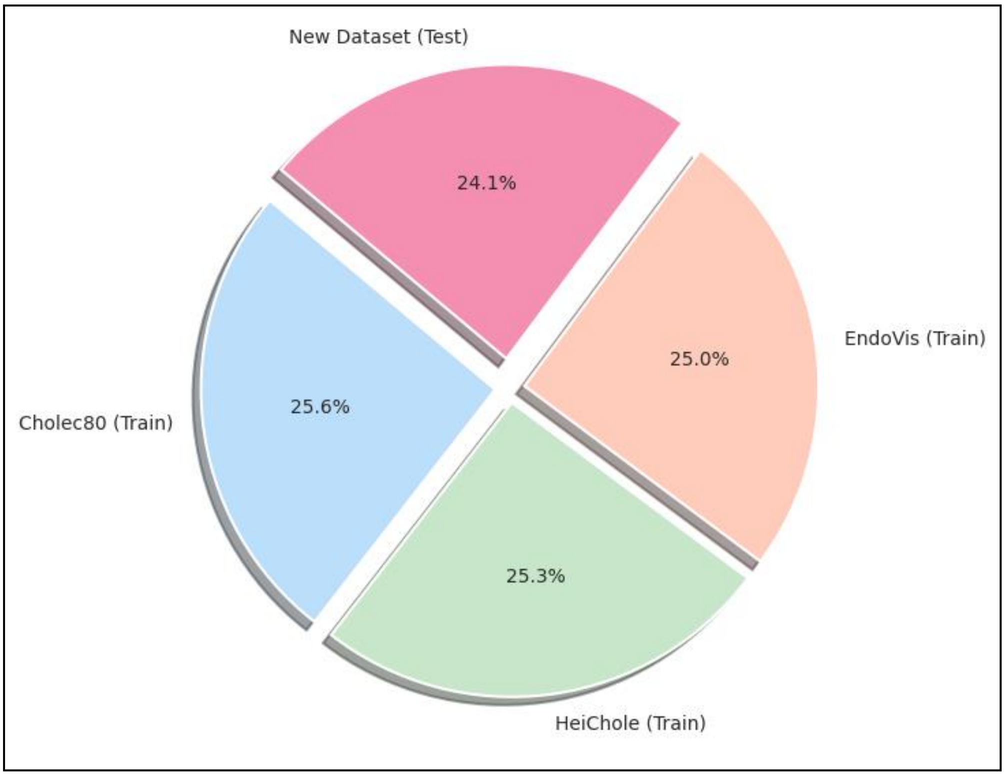

Figure 7. Pictorial Representation of Generalization Performance Across Different Surgical Datasets

Another significant challenge is interpretability and trust in AI-generated reports. While the model demonstrated high accuracy, medical professionals may still hesitate to rely entirely on automated documentation without human verification. Explainability techniques, such as attention maps and saliency visualizations, can help clinicians understand the decision-making process of the AI model and improve trust. Moreover, integrating human-in-the-loop mechanisms, where AI-generated reports are reviewed and corrected by surgeons, can enhance adoption and ensure clinical reliability. Ethical considerations and compliance with medical data privacy regulations must also be addressed. Ensuring that AI-driven documentation adheres to HIPAA and GDPR guidelines is essential to protect patient data. Implementing secure on-premises AI solutions or federated learning approaches can help mitigate privacy risks while maintaining model performance (As demonstrated in the above Figure 7). Bias detection frameworks should be incorporated to prevent any disparities in AI-generated reports based on patient demographics or surgical techniques. Our study demonstrates that hybrid vision-language models offer a highly accurate, efficient, and scalable solution for real-time surgical report generation. While challenges such as data scarcity, generalization, and interpretability remain, continuous advancements in self-supervised learning, model optimization, and AI transparency will further enhance the practical adoption of these systems in modern surgical environments. The integration of AI-driven documentation into smart operating rooms has the potential to revolutionize medical workflows, reduce administrative burdens, and improve overall patient care through standardized and real-time reporting.

6. CONCLUSION

Hybrid vision-language models (VLMs) represent a transformative advancement in surgical documentation, offering real-time, automated, and accurate report generation by integrating computer vision and natural language processing. Through the fusion of high-definition surgical video analysis and clinically coherent language generation, these systems significantly reduce the manual burden on clinicians, enhance the consistency and completeness of documentation, and improve interoperability with electronic health records. The experimental evaluation of the proposed model highlights its high accuracy in detecting surgical components, strong alignment with human-written reports, and efficient real-time performance, affirming its potential for clinical deployment. Despite challenges related to data diversity, annotation demands, generalization across surgical domains, and model interpretability, continuous innovations in self-supervised learning, domain adaptation, and ethical AI design are paving the way for more robust, scalable, and trustworthy implementations. As hybrid VLMs mature, their seamless integration into smart operating rooms will not only streamline surgical workflows but also elevate standards in patient care, safety, and medico-legal accountability.

REFERENCES

[1] Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Ramachandran, R. P., and Rasool, G. (2023). Transformers in time-series analysis: a tutorial. Circ. Syst. Sign. Process. 42, 7433–7466.

[2] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., et al. (2023a). Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond. arXiv Preprint arXiv:2308.12966.

[3] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. 3::23.

[4] Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., et al. (2021). “Scaling up visual and vision-language representation learning with noisy text supervision,” in International Conference on Machine Learning, Vol. 139, 4904–4916.

[5] Li, Z.; Zhao, W.; Du, X.; Zhou, G.; Zhang, S. Cross-modal retrieval and semantic refinement for remote sensing image captioning. Remote Sens. 2024, 16, 196.

[6] Mao, C.; Hu, J. ProGEO: Generating Prompts through Image-Text Contrastive Learning for Visual Geo-localization. arXiv 2024, arXiv:2406.01906.

[7] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., et al. (2023). Mistral 7B. arXiv Preprint arXiv:2310.06825.

[8] Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. Geochat: Grounded large vision-language model for remote sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27831–27840.

[9] Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and A Large Vision-Language Model for Remote Sensing. arXiv 2024, arXiv:2306.11300. [Google Scholar] [CrossRef]

[10] Kim, J.-H., Jun, J., and Zhang, B.-T. (2018). Bilinear attention networks. Adv. Neural Inform. Process. Syst. 31, 1564–1574.

[11] He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. (2020). PathVQA: 30000+ questions for medical visual question answering. arXiv Preprint arXiv:2003.10286.

[12] Bigolin Lanfredi R., Zhang M., Auffermann W. F., Chan J., Duong P.-A. T., Srikumar V., et al. (2022). Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Sci. Data 9:1441. 10.1038/s41597-022-01441-z

[13] Dou, Z.-Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., et al. (2022). “An empirical study of training end-to-end vision-and-language transformers,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (New Orleans, LA), 18145–18155.

[14] Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med. 30, 850–862.

[15] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: a visual language model for few-shot learning. Adv. Neural Inform. Process. Syst. 35, 23716–23736.