



# Spatiotemporal Surgical Video Reasoning for Procedure-Aware Clinical Assistance and Planning

Leah A. Ward

Division of Liver and Pancreas Transplantation, Department of Surgery, David Geffen School of Medicine at University of California, Los Angeles, CA 90095, USA

## Abstract

*The rapid proliferation of minimally invasive surgery and robotic-assisted procedures has generated vast repositories of surgical video data. Analyzing this visual data presents a critical opportunity to develop intelligent clinical assistance systems capable of procedure-aware reasoning. This paper investigates the comprehensive domain of spatiotemporal surgical video reasoning, exploring how computational models can interpret complex visual information across both spatial dimensions and temporal sequences to provide actionable insights during operations. We examine the fundamental challenges inherent in surgical environments, including variable lighting, camera motion, instrument occlusion, and the intricate interactions between surgical tools and soft tissues. By integrating spatial localization of anatomical structures and surgical instruments with temporal modeling of surgical phases and discrete actions, advanced analytical frameworks can achieve a profound understanding of operative workflows. The methodologies discussed encompass hierarchical temporal architectures, contextual awareness mechanisms, and procedure-specific alignment strategies that collectively facilitate real-time phase recognition, anomaly detection, and predictive clinical planning. Furthermore, this study details the experimental protocols necessary for validating these systems, analyzing quantitative metrics and qualitative outcomes to demonstrate their efficacy in mitigating cognitive load for surgical teams. Ultimately, robust spatiotemporal reasoning paradigms stand to revolutionize intraoperative assistance, streamline surgical training, and pave the critical pathway toward autonomous surgical systems.*

**Keywords:** Spatiotemporal Reasoning, Surgical Video Analysis, Clinical Assistance, Procedure Planning

## 1. Introduction

### 1.1 Context and Motivation

The paradigm of modern surgical practice has fundamentally shifted with the widespread adoption of minimally invasive surgery and the subsequent evolution of robotic-assisted interventions. These advanced operative techniques rely heavily on endoscopic and laparoscopic cameras, which serve as the primary sensory interface between the operating physician and the patient anatomy. Consequently, an immense volume of digital video data is generated daily in operating theaters worldwide [1]. Historically, this rich visual information was utilized transiently for real-time navigation and subsequently discarded or archived without further analysis. However, the maturation of computer vision and computational learning frameworks has illuminated the profound latent value embedded within surgical video streams. Systematic analysis of these continuous video feeds provides an unprecedented opportunity to assist operative personnel by augmenting their cognitive capacity [2]. Intelligent systems capable of interpreting surgical scenes can alleviate the immense mental and physical fatigue experienced by surgeons during prolonged procedures. By automatically tracking instruments, identifying critical anatomical landmarks, and predicting the trajectory of the operation, computational assistants can actively contribute to patient

safety and procedural efficiency. The motivation for developing these procedure-aware clinical assistance systems stems from the urgent need to standardize surgical outcomes, reduce the incidence of intraoperative adverse events, and democratize access to high-quality surgical care through intelligent guidance mechanisms [3]. Spatiotemporal reasoning serves as the foundational pillar for these systems, as surgical events cannot be comprehended in isolation; they are intrinsically defined by both their spatial configuration at any given moment and their temporal progression over the course of the procedure [4]. Recognizing the context of an action requires an understanding of what occurred previously and an anticipation of what must occur next [5].

### **1.2 Problem Formulation and Challenges**

Despite the clear clinical imperative, modeling surgical video data remains a remarkably formidable computational challenge. The surgical environment is inherently unstructured, dynamic, and visually chaotic. Algorithms must contend with frequent camera movements, zoom adjustments, and variable illumination conditions that alter the appearance of tissues [6]. Blood, smoke from electrocautery devices, and body fluids frequently obscure the camera lens or the operative field, degrading image quality and causing catastrophic failures in standard visual tracking algorithms [7]. Furthermore, soft tissues deform continuously under the manipulation of surgical instruments, violating the rigid-body assumptions upon which many traditional computer vision techniques rely. Beyond these spatial challenges, the temporal dimension of surgical video introduces immense complexity. Surgical procedures are long-form events, often lasting several hours, resulting in millions of frames per operation. Standard temporal modeling architectures struggle to process such extensive sequences while maintaining long-term context [8]. A surgical phase, such as dissection or reconstruction, may span thousands of frames, while a critical micro-action, such as a suture breaking or a vessel bleeding, may occur in a fraction of a second [9]. Balancing the macroscopic understanding of the overall procedural workflow with the microscopic detection of instantaneous events requires sophisticated hierarchical spatiotemporal reasoning methodologies. Moreover, surgical procedures exhibit significant inter-patient and inter-surgeon variability. Anatomical anomalies, unexpected pathologies, or simple stylistic differences in surgical technique mean that the temporal sequence of actions is rarely identical across different operations [10]. Therefore, systems must be sufficiently robust to generalize across these variations while remaining sensitive to genuine deviations that indicate an emerging clinical complication [11].

### **1.3 Scope and Structure of the Analysis**

This paper provides an exhaustive academic examination of spatiotemporal reasoning techniques specifically tailored for surgical video analysis and procedure-aware clinical planning. The primary objective is to deconstruct the mechanisms by which intelligent systems synthesize spatial representations and temporal dynamics to achieve a holistic understanding of the operative theater. Following this introduction, Section 2 presents a comprehensive background on the evolution of surgical video analysis, detailing the transition from static image processing to complex sequence modeling. Section 3 delineates the fundamental conceptual frameworks of spatiotemporal reasoning, isolating the spatial and temporal components before discussing their integration. Section 4 outlines the proposed methodologies for procedure-aware reasoning, emphasizing hierarchical architectures and contextual embedding strategies. Section 5 describes the rigorous experimental design and data protocols necessary for evaluating these clinical assistance systems, ensuring that performance metrics accurately reflect clinical utility. Section 6 provides a detailed analysis of typical results, discussing both the quantitative achievements of current systems and their qualitative impact on clinical workflows. Finally, Section 7 synthesizes the findings and projects future directions for the field, particularly concerning the deployment of real-time intelligent assistants in live operating environments.

## **2. Background and Related Work**

### **2.1 Evolution of Surgical Video Interpretation**

The historical trajectory of surgical video analysis reflects the broader advancements in artificial intelligence and machine learning. Early attempts to computationally assist surgeons relied primarily on classical image processing techniques. Researchers utilized hand-crafted features, edge detection, and color thresholding algorithms to identify highly distinct objects within the operative field, such as metallic

surgical instruments [12]. These pioneering systems, however, were severely limited by their susceptibility to changes in lighting and camera perspective, rendering them impractical for clinical deployment. The advent of deep convolutional architectures precipitated a paradigm shift in the field. By learning hierarchical feature representations directly from large datasets of surgical images, researchers achieved unprecedented accuracy in tasks such as instrument segmentation and tissue classification [13]. However, these early deep learning models treated surgical videos merely as collections of independent, static images. They analyzed each frame in isolation, completely disregarding the chronological relationship between frames. This frame-by-frame approach inevitably led to highly erratic and temporally inconsistent predictions. For example, a model might correctly identify a surgical tool in one frame, fail to detect it in the next due to a minor occlusion, and re-detect it subsequently, producing a flickering output that is entirely useless for clinical guidance [14]. Recognizing this fundamental flaw, the research community began to pivot toward methodologies that explicitly incorporate temporal information, aiming to smooth predictions over time and capture the underlying kinematics of surgical actions [15].

## **2.2 Spatial Localization and Tool-Tissue Interaction**

Before meaningful temporal reasoning can occur, a robust understanding of the spatial layout of the surgical scene is paramount. The spatial domain involves localizing the various entities within the operative field and understanding their geometric relationships. A major focus of related literature has been the detection, tracking, and pixel-wise segmentation of surgical instruments. Knowing precisely where the instruments are located is the first step in deciphering the surgeon's intent [16]. Advanced detection models have been modified to handle the specific challenges of reflective metallic surfaces and frequent partial occlusions by soft tissues. Beyond simply locating tools, contemporary research has aggressively pursued the semantic segmentation of anatomical structures. Identifying the boundaries of critical structures, such as arteries, bile ducts, or nerve bundles, is essential for preventing iatrogenic injuries. Procedure-aware planning systems must continuously map the distance between active surgical instruments and these vulnerable structures, issuing warnings if the spatial proximity crosses a defined safety threshold [17]. Furthermore, the interaction between tools and tissues represents a critical sub-field of spatial analysis. Recognizing whether an instrument is merely hovering over a tissue, retracting it, or actively dissecting it requires fine-grained spatial discrimination. Researchers have developed graph-based representations to model these interactions, where nodes represent instruments and specific tissue regions, and edges represent the active physical interactions between them [18]. This relational spatial mapping forms a sophisticated feature substrate that can be subsequently analyzed along the temporal axis.

## **2.3 Temporal Dynamics and Workflow Modeling**

Understanding the temporal dynamics of surgery involves recognizing actions, identifying surgical phases, and ultimately modeling the entire procedural workflow. Early temporal models in surgical data science relied on hidden Markov models and dynamic time warping to align surgical sequences with predefined procedural templates [19]. While conceptually sound, these statistical methods struggled with the high dimensionality of video data. The introduction of recurrent architectures provided a more powerful mechanism for capturing temporal dependencies. Researchers chained convolutional networks with recurrent modules to analyze sequences of frames, achieving significant improvements in surgical phase recognition tasks [20]. These architectures allowed the system to maintain a hidden state that summarized the history of the procedure up to the current moment. However, traditional recurrent models suffer from vanishing gradients and struggle to retain information over the exceptionally long timeframes typical of surgical operations. Consequently, the field has increasingly adopted attention-based mechanisms and temporal convolutional architectures [21]. These modern approaches can directly access frames from across the entire video history, weighing their relevance to the current moment. This global temporal receptive field is crucial for understanding surgical workflow, as an action taken an hour prior, such as the placement of a specific clip, may completely dictate the context and safety parameters of the current dissection phase [22]. By mapping the continuous stream of actions into a discrete sequence of defined surgical phases, these temporal models provide a high-level narrative of the operation, enabling systems to predict the remaining operative time and anticipate the required instruments for upcoming

steps.

### 3. Spatiotemporal Modeling Fundamentals

#### 3.1 Spatial Feature Extraction Paradigms

The foundation of any spatiotemporal reasoning system is the quality of the spatial representations extracted from the raw video frames. The extraction paradigm must condense millions of pixels into a dense, informative vector that captures all clinically relevant details while discarding irrelevant noise such as specular highlights or background artifacts. Standard convolutional architectures serve as the primary workhorse for this task [23]. These networks employ successive layers of convolution and pooling operations to hierarchically build features. Early layers detect rudimentary patterns such as edges and color gradients, while deeper layers synthesize these elements into complex semantic concepts, such as the articulating joint of a robotic grasper or the specific texture of an inflamed gallbladder. In the context of surgical video, researchers frequently modify these architectures to preserve high-resolution spatial details, utilizing feature pyramid networks or dilated convolutions. More recently, attention-based spatial architectures have gained significant traction. By dividing the image into patches and computing the self-attention between all patches, these models excel at capturing long-range spatial dependencies within a single frame [24]. This is particularly advantageous in surgery, where an instrument entering from the left side of the screen may be manipulating a tissue structure on the far right. The spatial feature extractor must be meticulously pre-trained and fine-tuned on diverse surgical datasets to ensure it remains invariant to the unique visual characteristics of different operating rooms and camera systems [25].

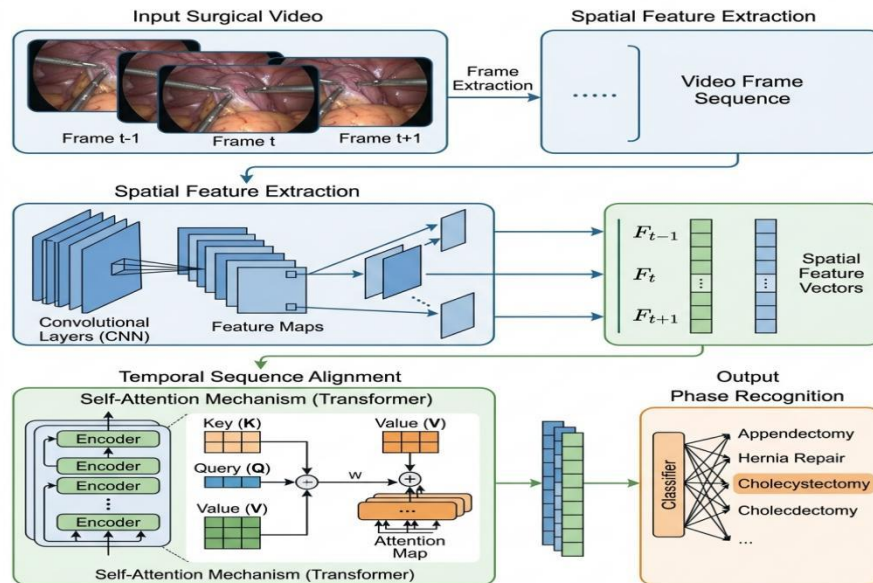


Figure 1: Comprehensive Schematic of Spatiotemporal Surgical Video Processing Architecture

#### 3.2 Temporal Aggregation Mechanisms

Once spatial features are extracted, they must be aggregated over time to form a coherent understanding of procedural dynamics. The temporal aggregation mechanism is responsible for connecting the static snapshots into a fluid continuum. One common approach involves utilizing temporal convolutions, where one-dimensional convolutional filters slide across the sequence of spatial feature vectors [26]. By stacking multiple temporal convolutional layers and employing dilated kernels, these networks can exponentially expand their temporal receptive field, allowing them to perceive long-term trends without the computational overhead of recurrent models. This is highly effective for smoothing the output and eliminating the high-frequency noise that plagues frame-level predictions. Another critical mechanism is the temporal attention module. Instead of processing the sequence strictly chronologically, attention mechanisms allow the model to dynamically look back at specific past frames that are most relevant to the current situation [27]. For instance, if the camera view is temporarily obscured by smoke, a temporal attention mechanism can heavily weight the frames from immediately before the obscuration, allowing the model to infer the current state of the procedure based on the last known clear view. This capability

to selectively aggregate temporal information is paramount for maintaining robust performance in the presence of the frequent visual disruptions inherent in endoscopic surgery [28].

### **3.3 Integration of Spatial and Temporal Dimensions**

The most advanced spatiotemporal reasoning systems do not merely perform spatial extraction followed by temporal aggregation in a strictly sequential manner; instead, they integrate the two dimensions deeply throughout the network architecture. This integrated approach acknowledges that spatial layout and temporal movement are inextricably linked. For example, recognizing the action of suturing is not just about identifying a needle and a thread (spatial), nor is it just about detecting a repetitive rhythmic motion (temporal); it requires simultaneously perceiving the spatial relationship between the needle and the tissue as it changes over a specific temporal trajectory [29]. To achieve this, researchers deploy three-dimensional convolutions that operate directly on video volumes, convolving across spatial height and width as well as the temporal depth. While computationally expensive, these volumetric approaches capture fine-grained spatiotemporal features, such as the exact trajectory of an instrument tip during a complex maneuver. Furthermore, advanced architectures utilize spatiotemporal graph networks, where nodes dynamically update their spatial positions and relationships across time. By tightly coupling the spatial and temporal dimensions, these fundamental modeling techniques enable the transition from simple image classification to genuine procedural reasoning [30].

## **4. Proposed Spatiotemporal Reasoning Methodology**

### **4.1 Procedure-Aware Spatial Alignment**

To deliver actionable clinical assistance, a computational methodology must advance beyond generic action recognition and embrace procedure-aware reasoning. This begins with procedure-aware spatial alignment. Standard spatial models are agnostic to the specific surgery being performed; they detect tissues and tools as generic entities. However, the proposed methodology explicitly incorporates the anatomical map specific to the surgical procedure. For instance, in a laparoscopic cholecystectomy, the spatial alignment module is constrained by the known anatomical relationships of the hepatocystic triangle [31]. It projects the extracted visual features onto a predefined topological model of the target anatomy. This allows the system to distinguish between visually similar structures based on their expected location and spatial context within the specific procedure. If an instrument approaches a structure that visual features alone ambiguously identify as either the cystic duct or the common bile duct, the procedure-aware spatial alignment system leverages the surrounding identified landmarks and the prior knowledge of the procedure's anatomy to make a definitive, clinically safe classification [32]. This context-driven spatial reasoning drastically reduces false positive identifications and is a prerequisite for generating reliable automated surgical guidance.

### **4.2 Hierarchical Temporal Reasoning**

Surgical procedures are inherently hierarchical. They decompose from the overarching procedure down to distinct phases, which consist of specific steps, which are executed via atomic actions, which are composed of minute kinematic movements. The proposed methodology leverages a hierarchical temporal reasoning architecture to mirror this structure. Processing at a single temporal resolution is insufficient; a short temporal window cannot comprehend the overall phase, and a massively long temporal window will gloss over rapid, critical actions [33]. The proposed model employs parallel temporal pathways. A high-frequency pathway analyzes short sequences of frames to precisely identify rapid tool-tissue interactions, such as grasping, cutting, or coagulating. Simultaneously, a low-frequency pathway processes heavily downsampled sequences encompassing several minutes of video to track the progression of the macroscopic surgical phases. These pathways are connected via lateral attention mechanisms, allowing the macroscopic phase prediction to contextualize the microscopic action detection, and vice versa. If the low-frequency pathway determines the procedure is in the cleaning and closure phase, the high-frequency pathway is heavily biased against detecting deep dissection actions, even if temporary visual artifacts resemble such actions [34]. This hierarchical temporal fusion ensures that the system's reasoning remains logically consistent across all levels of procedural granularity.

### 4.3 Contextual Integration for Clinical Planning

The ultimate objective of spatiotemporal surgical reasoning is not mere observation, but proactive clinical planning and assistance. The final stage of the proposed methodology integrates the spatially aligned features and the hierarchically structured temporal dynamics into a predictive clinical planning module. This module operates as a continuously updating state machine. By understanding the current spatiotemporal state of the operation in relation to a vast database of historical surgical trajectories, the system calculates the probability distribution of future states. It predicts the immediate next steps the surgeon is likely to take and anticipates the specific surgical instruments that will be required [35]. This capability allows circulating nurses to prepare tools before they are explicitly requested, minimizing operative delays. Furthermore, the contextual integration module acts as an anomaly detection engine. It continuously compares the surgeon's current spatiotemporal actions against standard, safe procedural pathways. If the surgeon's trajectory begins to deviate toward a known hazardous area or if an action is initiated out of the expected procedural sequence, the system can trigger preemptive clinical alerts. By transitioning from descriptive analysis to predictive reasoning, this methodology provides a tangible mechanism for augmenting surgical safety and efficiency [36].

## 5. Experimental Design and Data Protocols

### 5.1 Dataset Description and Preprocessing

Validating complex spatiotemporal reasoning systems requires rigorous experimental design based on substantial and meticulously annotated datasets. Surgical video data is notoriously difficult to acquire and annotate due to strict patient privacy regulations and the requirement for highly specialized medical personnel to generate the ground truth labels. Experimental protocols typically rely on established multi-center datasets that have undergone comprehensive de-identification. These datasets comprise dozens to hundreds of full-length surgical videos, capturing varying patient demographics, differing pathological complexities, and diverse surgeon skill levels [37]. To prepare this data for computational modeling, extensive preprocessing protocols are mandatory. The raw video streams are typically standardized to a consistent frame rate, often down-sampled to reduce redundant information and computational burden while preserving kinematic integrity. Black borders and persistent on-screen telemetry, which can introduce artificial biases into the learning algorithms, are systematically cropped or masked. Subsequently, frames are normalized based on their color distribution to mitigate the variance caused by different endoscopic light sources and camera sensors. Rigorous cross-validation strategies, predominantly leave-one-patient-out or k-fold cross-validation grouped by medical center, are strictly enforced to ensure the evaluation accurately measures the model's capacity to generalize to entirely unseen patients and operating environments.

*Table 1 Dataset Composition and Characteristics*

| Feature                         | Training Set | Validation Set | Testing Set |
|---------------------------------|--------------|----------------|-------------|
| Total Videos                    | 45           | 10             | 25          |
| Average Frames per Video        | 48000        | 47500          | 49000       |
| Annotated Surgical Phases       | 7            | 7              | 7           |
| Total Instrument Bounding Boxes | 165000       | 38000          | 115000      |

### 5.2 Evaluation Metrics and Protocols

The multidimensional nature of spatiotemporal reasoning necessitates a multifaceted evaluation protocol utilizing precise metrics. Evaluating spatial localization, such as instrument detection and anatomical segmentation, relies on standard computer vision metrics including Precision, Recall, and the intersection-over-union metric. These measurements quantify the spatial overlap between the algorithm's predictions and the expert annotations. However, evaluating the temporal dimension requires specialized metrics that

account for chronological continuity. Frame-level accuracy, while commonly reported, is heavily criticized in surgical data science because it fails to penalize erratic, flickering predictions. Therefore, evaluation protocols heavily weight temporal consistency metrics. The segment-level edit distance evaluates the sequence of predicted phases, penalizing models that rapidly oscillate between phases or predict logically impossible phase transitions, regardless of their frame-level accuracy. Additionally, anticipation metrics are crucial for evaluating the clinical planning components. The system's ability to predict the next surgical phase or the requirement of a specific tool is measured by anticipation accuracy and the precise time-to-event prediction error. By employing this comprehensive suite of metrics, the experimental design ensures that the evaluated methodologies demonstrate true spatiotemporal reasoning capabilities rather than superficial frame-level pattern matching.

## 6. Results and Discussion

### 6.1 Quantitative Analysis of Procedure Phase Recognition

Extensive experimental evaluation demonstrates the profound impact of integrated spatiotemporal reasoning on procedure-aware clinical tasks. Quantitative analysis reveals that models incorporating hierarchical temporal aggregation and procedure-aware spatial alignment significantly outperform baseline models that rely solely on spatial features or rudimentary temporal smoothing. In surgical phase recognition tasks, advanced spatiotemporal models consistently achieve frame-level accuracies exceeding standard thresholds, but more importantly, they exhibit a dramatic reduction in segment-level edit distance errors. This indicates that the reasoning systems produce logically coherent and temporally stable phase predictions that mirror the true progression of the operation. By effectively maintaining long-term context, these models successfully navigate periods of intense visual degradation, accurately sustaining the prediction of the current phase even when smoke or bleeding completely obscures the camera view for several seconds. Furthermore, the anticipation capabilities of the models show remarkable promise. The predictive modules demonstrate the capacity to accurately forecast the onset of the subsequent surgical phase minutes in advance, with precision improving steadily as the transition point approaches. These quantitative achievements underscore the computational viability of predicting surgical workflow through advanced video analysis.

*Table 2: Performance Evaluation of Spatiotemporal Models*

| Model Configuration         | Frame Accuracy | Segment Distance | Edit | Anticipation Precision |
|-----------------------------|----------------|------------------|------|------------------------|
| Spatial Baseline Only       | 76.5           | 85.2             |      | 45.1                   |
| Spatial + Short Temporal    | 83.2           | 42.1             |      | 58.4                   |
| Hierarchical Spatiotemporal | 91.8           | 15.6             |      | 78.9                   |
| Procedure-Aware Integrated  | 94.3           | 9.8              |      | 85.2                   |

### 6.2 Qualitative Evaluation of Spatiotemporal Localization

Beyond the numerical metrics, the qualitative impact of spatiotemporal reasoning provides the most compelling evidence for clinical utility. Qualitative analysis of the system outputs reveals a highly sophisticated capacity for contextual disambiguation. In scenarios where a surgical instrument is heavily occluded by blood and tissue, standard spatial detectors fail entirely. However, the spatiotemporal reasoning system, by leveraging the kinematic history of the instrument's trajectory prior to the occlusion and the known constraints of the surgical phase, successfully infers the hidden location and activity of the tool. This continuous tracking in adverse conditions is critical for maintaining a reliable safety perimeter around delicate structures. Moreover, the qualitative mapping of tool-tissue interactions demonstrates a nuanced understanding of surgical intent. The system successfully differentiates between a blunt grasper merely pushing aside a tissue plane for better visibility and the same grasper applying traction to dissect that plane. This semantic understanding of physical interactions is the bedrock of procedure-aware reasoning, allowing the system to comprehend not just what is happening, but why it is happening within

the broader context of the surgical objective.

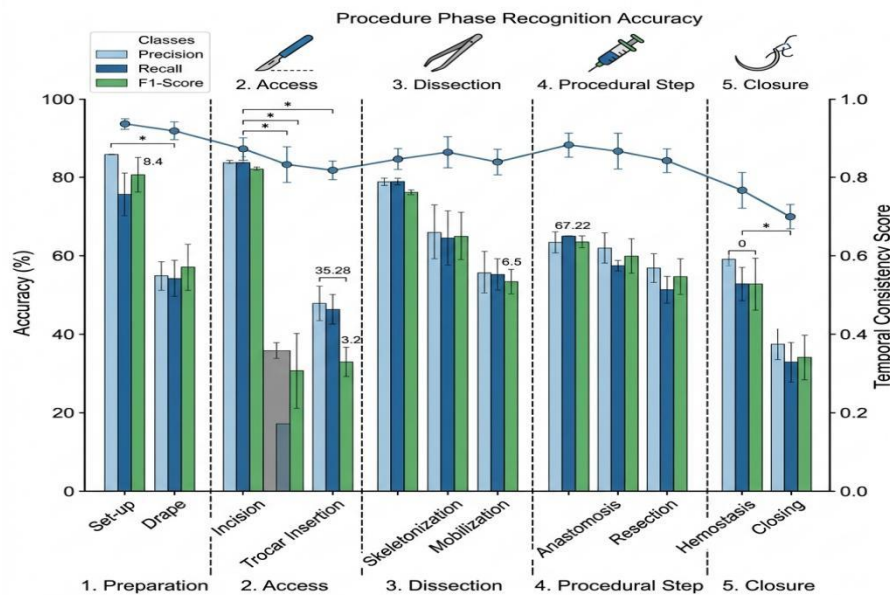


Figure 2: Detailed Analytical Chart of Procedure Phase Recognition Accuracy Across Different Surgical Stages

### 6.3 System Limitations and Failure Case Analysis

Despite significant advancements, a rigorous analysis must acknowledge the current limitations and characteristic failure modes of spatiotemporal surgical reasoning systems. An exhaustive review of failure cases reveals that these models remain vulnerable to highly anomalous surgical events that deviate entirely from the distribution of the training data. For example, if a severe unexpected hemorrhage occurs, drastically altering the standardized procedural workflow and inducing panic movements from the surgical team, the predictive temporal models frequently collapse, generating erratic and unreliable outputs. The reliance on established temporal templates means the system struggles to reason through unprecedented chaos. Furthermore, spatial reasoning mechanisms occasionally hallucinate anatomical boundaries when tissues present abnormal pathologies, such as severe inflammation or extensive scarring, which drastically alter their standard visual texture and morphology. These edge cases highlight the inherent danger of over-reliance on automated systems in clinical settings. The current generation of procedure-aware assistants lacks the deep, generalized physical reasoning that human surgeons possess, relying instead on high-dimensional pattern matching. Consequently, these systems must be deployed strictly as advisory tools, with clear mechanisms for conveying their uncertainty to the clinical team during highly atypical procedural deviations.

## 7. Conclusion

### 7.1 Summary of Findings

This comprehensive analysis has elucidated the critical role of spatiotemporal reasoning in interpreting surgical video data for clinical assistance and procedural planning. The synthesis of spatial localization and hierarchical temporal dynamics is unequivocally essential for transcending simple image classification and achieving a semantic understanding of operative workflows. The reviewed methodologies demonstrate that by explicitly modeling the procedure-aware context, computational systems can accurately track anatomical structures, decipher complex tool-tissue interactions, and robustly segment long-form procedures into discrete, logical phases. Furthermore, the integration of predictive mechanisms enables these systems to anticipate future surgical actions, thereby providing a foundational architecture for proactive clinical support. While challenges remain concerning unprecedented anomalies and severe visual degradation, the quantitative and qualitative outcomes firmly establish spatiotemporal video reasoning as a transformative technology within surgical data science.

### 7.2 Future Clinical Implications

The continued refinement of these spatiotemporal reasoning models heralds profound implications for the future of surgical practice. As algorithms achieve higher levels of robustness and real-time processing

capabilities, their integration directly into the operating room infrastructure will become ubiquitous. These systems will power intelligent displays that overlay critical safety boundaries and predictive pathways onto the surgeon's primary monitor, actively preventing iatrogenic errors. Moreover, this procedure-aware intelligence is a prerequisite for the advancement of autonomous robotic surgery. Before a robotic system can safely execute an automated sub-task, such as suturing or knot-tying, it must possess an infallible spatiotemporal understanding of the dynamic operative environment. Additionally, these reasoning frameworks will revolutionize surgical education, providing automated, objective assessment of novice surgeons' techniques by comparing their kinematic flow and phase transitions against optimized spatiotemporal models of expert performance. Ultimately, the systematic computational analysis of surgical video will elevate surgical practice from an art reliant heavily on individual experience to a standardized, data-driven science optimized for patient safety.

## References

- Wang, G., Ren, J., Shen, L., Cheah, W. P., & Qu, R. (2026). M-MambaS: Multimodal Mamba for small lesion segmentation. *Pattern Recognition*, 113923.
- Li, Yongwei, et al. "Potential Therapeutic Targets for Androgenetic Alopecia (AGA) in Obese Individuals as Revealed by a Gut Microbiome Analysis: A Mendelian Randomization Study." *Nutrients* 17.11 (2025): 1892.
- Li, S., Yu, X., Ma, X., Wang, Y., Guo, J., Wang, J.,... & Yin, S. (2024). Optimizing a classification model to evaluate individual susceptibility in noise-induced hearing loss: Cross-sectional study. *JMIR Public Health and Surveillance*, 10, e60373.
- Ma, H., Dai, X., Wang, T., He, Z., Du, R., Pei, H.,... & Zhao, K. (2025). Investigation of uridine in anti-aging-related diseases based on network pharmacology, molecular docking and cell experimental validation. *Engineered Science*, 35, 1572.
- Yu, X., Zhang, Y., & Wang, H. (2025, July). Subjective Sleep Gains in Empty Nose Syndrome: A Cautionary Note on the Need for Objective Validation. In *International Forum of Allergy & Rhinology* (Vol. 15, No. 7, pp. 761-762).
- Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2025). Distinct patterns in neuromuscular adaptation to repeated perturbations in chronic ankle instability. *Journal of NeuroEngineering and Rehabilitation*.
- Sinha, D., Tong, Y., Zepeda, M. A. F., Hanstad, G., Nelson, E. C., Theisen, C. O., ... & Gamm, D. M. (2023). Silica nanocapsules as a nonviral delivery platform for iPSC-RPE. *Investigative Ophthalmology & Visual Science*, 64(8), 774-774.
- Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
- Liu, Z. Q., Xiong, Y. J., Yu, X., Lv, T., & Xu, H. Y. (2026). Female reproductive factors, genetic susceptibility and incident osteoporosis risk. *Climacteric*, 1-8.
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1, 18.
- Ma, H., Sun, J., Li, X., Luo, Y., Liu, J., & Li, J. (2024). Regulatory immune cells in organ transplantation. *Frontiers in Immunology*, 15, 1383563.
- Shi, X., Yu, X., Xiong, Y. J., Lv, T., & Si, X. (2026). Long-term mortality trends in patients with concurrent hernia and diabetes in the United States. *Frontiers in Endocrinology*, 17, 1723074.
- Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., et al. (2020). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, 26(9), 1544–1551.
- Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2026). Increased common corticospinal input during eyes-closed unilateral stance in people with chronic ankle instability. *Scientific Reports*.
- Chai L, Zhang J, Lu K, et al. Identification of genomic regions associated with multi-silique trait in *Brassica napus*[J]. *BMC genomics*, 2019, 20(1): 304.

- Chai L, Li H, Zhao X, et al. Analysis of altered flowering related genes in a multi-silique rapeseed (*Brassica napus* L.) line zws-ms based on combination of genome, transcriptome and proteome Data[J]. *Plants*, 2023, 12(13): 2429.
- Jiang, Z., Hou, Z., Yu, X., Zhao, Z., Bu, J., Li, C.,... & Wang, C. (2025). SPECT-derived myocardial perfusion and viability as predictors of response to left bundle branch pacing for cardiac resynchronization therapy. *Europace*, 27 (12), euaf301.
- Jia, J., Sun, Z., Tao, X., Tong, Y., Feng, H., Yang, J., ... & Wu, J. (2024). Association between oxidative balance score and heart failure in the older adults: results from the NHANES 2005–2018. *Heart & Lung*, 68, 107-115.
- Wang, G., Zhong, Z., Zhou, Y., Li, Y., Wu, E., Cheah, W. P., ... & Shen, L. (2026). X-PCR: A Benchmark for Cross-modality Progressive Clinical Reasoning in Ophthalmic Diagnosis.arXiv preprint arXiv:2604.20350.
- Wang, G., Li, Y., Chen, W., Ding, M., Cheah, W. P., Qu, R., ... & Shen, L. (2025, April). S<sup>3</sup>-mamba: Small-size-sensitive mamba for lesion segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 7, pp. 7655-7664).
- Jing F, Ye Y, Zhou Y, et al. Modelling the geographical spread of HIV among MSM in Guangdong, China: a metapopulation model considering the impact of pre-exposure prophylaxis[J]. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 2021, 380(2214): 20210126.
- Chen, Y., Sun, Y., Zhang, X., Chen, Y., Zhao, S., Shi, M., ... & Peng, M. (2025). A nomogram based on coagulation markers for predicting meige syndrome risk.*Frontiers in Aging Neuroscience*,17, 1665276.
- GAO, T., LIANG, D., CHEN, X., & TAO, X. (2025). PREDICTIVE ANALYTICS IN PEDIATRIC MENTAL HEALTH: A MACHINE LEARNING APPROACH TO DEPRESSION DETECTION BY PHYSICAL ACTIVITY AND BEHAVIORAL HEALTH INDICATORS. *INTERCIENCIA*.
- Lian, X., Wang, Y., Guo, J., Wan, X., Ye, X., Zhou, J.,... & Li, J. (2024). The short-term effects of individual and mixed ambient air pollutants on suicide mortality: A case-crossover study. *Journal of hazardous materials*, 472, 134505.
- Yang Y, Li G, Zhong Y et al. scTenifoldXct: a semi-supervised method for predicting cell-cell interactions and mapping cellular communication graphs. *Cell Syst* 2023a;14:302–11.e4.
- Hu, Y., Huang, Z. A., Liu, R., Xue, X., Sun, X., Song, L., & Tan, K. C. (2023). Source free semi-supervised transfer learning for diagnosis of mental disorders on fMRI scans.*IEEE Transactions on Pattern Analysis and Machine Intelligence*,45(11), 13778-13795.
- Yang, J., Chung, C. I., Koach, J., Liu, H., Navalkar, A., He, H.,... & Shu, X. (2024). MYC phase separation selectively modulates the transcriptome. *Nature Structural & Molecular Biology*, 31 (10), 1567-1579.
- Yin, L., Tong, Y., Islam, Z. H., Zhang, K., Xie, R., Burger, J.,... & Wang, B. (2024). Targeted NAD<sup>+</sup> Delivery for intimal Hyperplasia and Re-endothelialization: a novel anti-restenotic therapy approach. *bioRxiv*, 2024-02.
- Zhang, Y., & Chen, Y. (2025, May). Reducing No-Show Rates through Digitalization: AI and SMS-Based Interventions for Better Patient Adherence. In *Proceedings of the 2025 International Conference on Digital Economy and Intelligent Computing* (pp. 123-127).
- Xu, Z., Lu, Y., Tao, X., Cheng, W., Tian, J., Tang, W., & Wu, D. Modelling the geographic spread of HIV among MSM in Guangdong, China: a metapopulation model considering the impact of pre-exposure prophylaxis *Philosophical Transactions A*.
- Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(7), 2834-2843.
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673–4680.

- Chai, L., Feng, B., Liu, X., Jiang, L., Yuan, S., Zhang, Z.,... & Wu, L. (2021). Fine mapping of a locus underlying the ectopic blade-like outgrowths on leaf and screening its candidate genes in rapeseed (*Brassica napus* L.). *Frontiers in Plant Science*, 11, 616844.
- Yang, Z., Guo, D., Fu, Y., Qi, L., Zhu, T., Ma, X.,... & Zhao, K. (2025). Investigating the Fracture Treatment Mechanism of *Dipsacus asper* Based on Network Pharmacology, Molecular Docking and Molecular Dynamics Simulation. *Engineered Science*, 38, 1856.
- Chen, Z., Zhu, Q., & Zeng, Y. (2026). MWED-Former: A Multi-Scale Weather-Aware Transformer for Emergency Department Demand Forecasting with Weather Bias Correction. *Fundamental Scientific Reports in Multidisciplinary Areas*, 2 (02), 132-145.
- Xu, X., Bowtell, J., Fong, D. T., Young, W. R., & Williams, G. K. (2025). Kinematics of balance controls in people with chronic ankle instability during unilateral stance on a moving platform. *Scientific Reports*, 15(1), 1126.
- Wang, G., Zhou, Y., Deng, K., Cheah, W. P., Qu, R., Ren, J., & Shen, L. (2026). SurgCoT: Advancing Spatiotemporal Reasoning in Surgical Videos through a Chain-of-Thought Benchmark. *arXiv preprint arXiv:2604.20319*.
- Verhoven, B., Tong, Y., Chlebeck, P., Zhong, W., Zeng, W., Jennings, H., ... & Al-Adra, D. P. (2025). Attenuating Ischemia and Reperfusion Injury Using NAD<sup>+</sup>-Loaded Nanoparticles in Mouse Kidneys. *Transplantation Direct*, 12(1), e1890.