Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
https://irojournals.com/jucct/

ORION: A Hierarchical Surgical World Model for Real-Time, Multi-Agent Operating-Room Intelligence

Dev Mehta[1], Charulatha R.T.[2]

SRM Institute of Science and Technology, Chennai, India.

E-mail: [1]dm8817@srmist.edu.in, [2]charular@srmist.edu.in

Abstract

Modern operating suites generate dense, heterogeneous data streams endoscopic video, multi-parameter vitals, anesthesia registers, device telemetry, intraoperative imaging, and robotic kinematics yet the overwhelming majority of these signals are discarded, with the complete clinical reality of a procedure routinely collapsed into a single postoperative note. This documentation bottleneck loses roughly ninety-nine percent of the granular information produced during surgery, preventing systematic audit, objective skill assessment, and longitudinal research. We present ORION (Operating Room Intelligence and Optimization Network), an AI-native platform that transforms multi-sensory input into a continuously updated digital twin of the patient and operating environment. ORION is organised as a twelve-layer Hierarchical Surgical World Model (HSWM) spanning real-time perception, relational scene-graph construction, a Gaussian-splatting digital twin, temporal workflow modelling, predictive risk forecasting, knowledge grounding, and multi-agent clinical reasoning. A generative component, SurgWorld, mitigates kinematic-data scarcity by synthesising physically plausible surgical video and recovering pseudo-kinematics through an inverse-dynamics model. Because ORION is presented as a proposed architectural framework, the reported Cholec80 and AutoLaparo figures up to 94.6% and 89.5% accuracy at 238.7 FPS for the underlying state-space model are drawn from the constituent methods' published evaluations, not a new experiment; prospective validation of the integrated system is future work. The reporting layer

DOI: https://doi.org/10.36548/jucct.2026.3.008
Received: 06.06.2026, received in revised form: 09.07.2026, accepted: 01.08.2026, published: 12.08.2026

is designed to compress postoperative documentation to under thirty seconds. We further describe the low-latency execution architecture (NVIDIA Holoscan and Triton) required to stay below the 50 ms perceptual threshold, and map the platform onto India's CDSCO Software-as-a-Medical-Device regulatory pathway.

Keywords: Surgical Intelligence, Digital Twin, Surgical World Model, Multi-Agent Systems, Surgical Phase Recognition, Operating-Room Analytics.

1. Introduction

The modern operating suite is an environment of extreme data density coupled with cognitive fragmentation. During a single procedure, multiple clinicians coordinate their physical movements while interacting with a complex ecosystem of endoscopic feeds, multi-parameter vitals, anesthesia registers, device telemetry, intraoperative ultrasound, and robotic manipulator coordinates [1, 2]. Despite this abundance, surgical workflows remain analytical islands: the complete intraoperative record is typically reduced to a brief postoperative note, discarding the vast majority of the granular information generated during surgery and precluding systematic audit, objective training assessment, and longitudinal study.

1.1 The Information-Loss Problem

Because intraoperative signals are neither synchronised nor persisted, they cannot be queried after the fact. A surgeon who wishes to review the moment a clip failed, or a training director who wishes to compare a resident's instrument economy against a cohort, has no structured substrate to work from. The loss is not merely archival — it removes the evidentiary basis for the closed-loop feedback that has driven quality improvement in other high-reliability domains.

1.2 The ORION Platform

ORION is designed not as a passive recording utility but as an AI-native surgical intelligence platform that transforms multi-sensory and visual input into a continuously updated digital twin of the patient and the operating environment. It is structured as a Hierarchical Surgical World Model (HSWM) that transitions through distinct cognitive tiers — from real-time visual perception, through semantic scene understanding, predictive risk forecasting, and multimodal knowledge reasoning, to structured, automated clinical documentation.

1.3 Contributions

1. A twelve-layer HSWM that unifies perception, relational scene modelling, a volumetric digital twin, temporal workflow recognition, prediction, knowledge grounding, and multi-agent reasoning under one architecture.
2. An online phase-recognition model, SurgicalMamba, combining intensity-modulated temporal stepping with orthogonal state regramming to achieve linear per-frame inference at 238.7 FPS while improving accuracy over transformer and TCN baselines [6, 7].
3. A generative surgical world model, SurgWorld, trained by flow matching, that mitigates kinematic-data scarcity and enables pseudo-kinematic policy extraction through an inverse-dynamics model [11, 12].
4. A LangGraph-orchestrated multi-agent reasoning layer with cross-agent consensus scoring and mandatory human-in-the-loop override on safety-critical deviations [15, 16].
5. A deployment and compliance analysis covering sub-50 ms streaming execution on NVIDIA Holoscan/Triton and classification under India's CDSCO SaMD framework [21, 22].

2. Connected Operating-Suite Architecture

To resolve the fragmentation of intraoperative data, ORION acts as an integration gateway that aggregates and synchronises every available sensor stream in the operating suite, establishing a continuous, high-bandwidth pipeline as the foundation for real-time inference.

2.1 Architectural Taxonomy of Sensor Streams

Table 1 summarises the principal data-source categories, their telemetry profiles, ingestion protocols, and roles within the digital twin.

Table 1. Sensor-stream taxonomy ingested by the ORION integration gateway

Data-source category	Consumed telemetry / signal	Ingestion protocols	Role in the digital twin
Surgical personnel	Ambient positioning of 4–12 staff (scrub nurses, assistants, surgeons)	Ultra-wideband (UWB) tracking; multi-camera exocentric video	Workflow coordination, team ergonomics, resource bottlenecks
Visual sensors	4K/60 FPS endoscopic, laparoscopic, robotic camera feeds	SMPTE ST 2110, low-latency RTSP, PCIe capture	Primary feed for the perception engine
Instruments	Physical and active tool tracking across dozens of tools	Real-time CV detection, RFID tags	Tool transitions, usage frequencies, operational time
Physiological signals	Continuous multi-parameter vitals	HL7, serial interfaces, A/D streaming	Heart rate, SpO_2, BP integrated on the twin timeline
Anesthesia registry	Agent concentrations, IV drug records, ventilator data	IEEE 11073 SDC	Correlates physiology with surgical events
Device telemetry	Electrocautery power, insufflation, suction–irrigation	Modbus TCP, proprietary serial	Device metrics during cutting and dissection
Intraoperative imaging	Ultrasound sweeps, fluoroscopy, DICOM overlays	DICOM Q/R, RTSP, video loopback	Matches subsurface structures to surface models
Robotic kinematics	Joint Cartesian coordinates, arm configs, gripper apertures	High-speed UDP, ROS2	Ground-truth kinematics for skill profiling
Postoperative records	Operative notes, charts, pathology reports	HL7 FHIR	Grounds decisions in patient history

2.2 Technical Architecture Synthesis

Information flows from the hardware layer, through a real-time computer-vision cluster, into the multi-agent reasoning layer, and finally into the hospital's EHR infrastructure. Camera feeds enter a real-time AI gateway (SMPTE ST 2110 / RTSP), pass to a CV cluster (YOLO12, SAM2, nnU-Net, RT-DETR), and are assembled into a relational scene graph. The scene graph updates a Gaussian-splatting digital twin, which feeds a temporal state-space model (SurgicalMamba) for phase recognition. Verified recommendations from the multi-agent layer are then written to a knowledge graph and vector store, and back to the hospital dashboard and FHIR API.

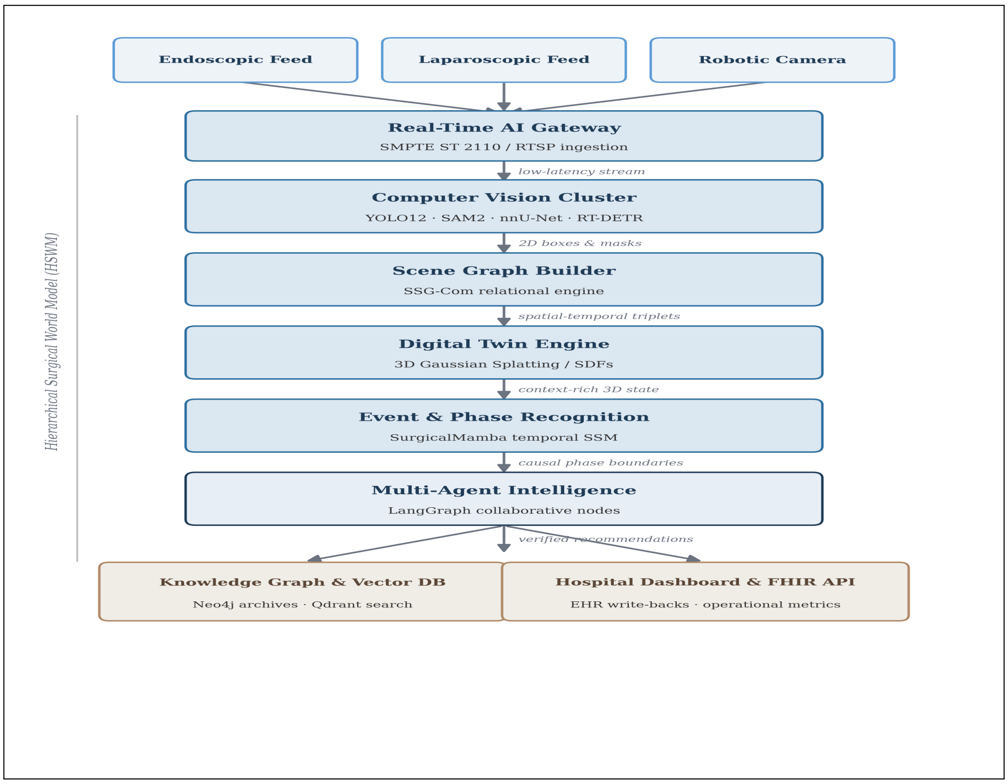

Figure 1. End-to-end architecture and data flow of the ORION platform

The twelve layers form a single directed pipeline in which each stage consumes a typed output from the stage above it and emits a representation at a higher level of abstraction, as illustrated in Figure 1. The perception layer (Layer 1) converts raw pixels into per-frame instrument detections and anatomical segmentation masks. These detections are consumed by the scene-graph layer (Layer 2), which encodes the active entities and their spatial and functional relations as subject-relation-object triplets. The triplet stream drives the digital-twin engine (Layer 3), which maintains a persistent volumetric reconstruction of the cavity and computes geometric quantities such as instrument-to-boundary distances that are not observable in any single frame.

The temporal layer (Layer 4) consumes the frame-level features and the evolving twin state to infer the current surgical phase, and its phase boundaries in turn gate the predictive layer (Layer 5), which forecasts remaining duration, complication probability, and bleeding trajectories. Layers 1 through 5 therefore constitute the perception-to-prediction path, producing the state variables that the higher layers reason over. The procedural knowledge graph (Layer 6) grounds these local observations in a multi-institutional archive, and the reinforcement-learning support layer (Layer 7) uses the same state to propose low-strain tool trajectories.

At the cognitive tier, the multi-agent layer (Layer 8) is the principal consumer of all lower-layer outputs: each specialist agent subscribes to the specific streams relevant to its sub-domain — the Vision and Anatomy agents to the perception masks, the Risk agent to the scene graph and bleeding estimates, the Guideline agent to the phase context — and a coordinator node aggregates their outputs into a consensus score with a human-in-the-loop override. The remaining layers act as consumers of the consolidated state rather than producers of new inference: the memory architecture (Layer 9) indexes completed procedures for similarity search, the reporting layer (Layer 10) serialises the state history into a FHIR record, the skill-assessment layer (Layer 11) derives kinematic competency metrics, and the analytics dashboard (Layer 12) aggregates cross-procedure statistics. Data thus flows strictly upward through the perception, cognition, and documentation tiers, while the digital twin of Layer 3 serves as the shared state store that every higher layer reads from, ensuring that all modules operate on one temporally consistent representation of the procedure.

3. The Surgical Perception Engine (Layer 1)

The foundation of the HSWM is the Surgical Perception Engine, which transforms raw pixels into semantically segmented representations of the surgical cavity. It combines object detection, anatomical segmentation, and environmental anomaly detection into a single real-time inference pipeline.

3.1 Instrument Detection and Spatial Characterisation

Instrument tracking must identify multiple identical-looking metallic tools under occlusion, fluid splatter, and camera motion. ORION uses optimised YOLO12 and RT-DETR architectures to generate real-time tool-tracking data [3, 9]. Table 2 lists representative instrument classes and their tracked attributes.

Table 2. Instrument classes tracked by the perception engine

Instrument class	Primary action	Bounding representation	Tracked semantic attributes
Scalpel	Sharp tissue division	2D box [x,y,w,h] + tip keypoint	Position, velocity, usage duration, active user
Forceps	Traction, manipulation	2D box + jaw-aperture angle	Jaw position, grasp state, contact pressure
Needle holder	Suture guidance	Dual keypoints (jaw base, tip)	Grasp confidence, rotational velocity, suture index
Trocar	Channel creation	Linear axis vector (angle, depth)	Insertion point, angulation stability
Camera	Feedback, illumination	View-frustum vector	Movement speed, drift, lens obstruction
Stapler	Tissue division/closure	Polyhedral box around anvil	Open/closed state, deployment, alignment margin
Suction	Fluid removal	Tip keypoint + audio monitor	Flow status, position vs. pools, activation intervals
Clip applier	Vessel ligation	Dynamic keypoint along jaws	Clip presence, closure level, placement success

3.2 Anatomical Segmentation, Tissue States, and Environmental Anomalies

In parallel, pixel-level anatomical boundaries are mapped to identify structures, track targets, and avoid critical zones using SAM2, nnU-Net, and Mask2Former [4, 18, 19]. Each structure carries a semantic state classification and an associated real-time anomaly metric for example, a smoke-density index φ for electrocautery smoke, or an estimated blood-loss rate EBL_t for localised bleeding each paired with an actionable recommendation. Table 3 summarises the tissue-state and anomaly model.

Table 3. Anatomical segmentation, tissue states, and environmental anomalies

Target structure	Semantic states	Environmental anomaly	Anomaly metric / suggestion
Liver	Healthy, inflamed, bleeding, scarred, necrotic, cancerous	Electrocautery smoke	Smoke-density index, visibility loss, evacuation/suction prompt
Gallbladder	Healthy, inflamed, necrotic, gangrenous, scarred	Localised bleeding	Estimated blood-loss rate, source coordinate, trend forecast
Kidney	Healthy, tumour-bearing, scarred, necrotic	Vessel perforation	Breach flag, arterial/venous spray tracking
Colon	Healthy, ischemic, perforated, inflamed, cancerous	Visual occlusion	Lens-obstruction %, water-flush/clean alert
Blood vessels	Patent, calcified, occluded, bleeding	Tissue deformation	Surface-normal deviation maps, strain/tear risk
Fat tissue	Healthy, ischemic, necrotic	Instrument slips	High-velocity tool deviation $\rightarrow$ acoustic/visual alert
Tumours	Solid, infiltrative, cystic, benign, malignant	Thermal spread	Colour-space shift estimating thermal-injury boundaries
Nerves	Intact, compressed, divided	Structural tension	Displacement models identifying traction boundaries

4. The Surgical Scene Graph (Layer 2)

In addition to isolated object segmentations, ORION structures the clinical environment into a spatial-temporal relational graph, translating geometric coordinates into semantic relationships [4, 5]. The Surgical Scene Graph (SSG) is produced by the SSG-Com architecture, trained on the Endoscapes-SG201 dataset. Each node is an active entity (instrument, anatomical structure, or surgeon hand) and each directed edge is a spatial or functional relation, yielding unified triplets:

$$\mathcal{T} = (Subject, Relation, Object) \tag{1}$$

For example, a dissection sequence is represented as a conjunction of triplets:

$$\begin{aligned} \mathcal{T} dissect \;=\; &(\; Forceps[operator\_right\_hand],\; holding,\; Blood\,Vessel\;) \\ &\wedge (\; Blood\,Vessel,\; next\_to,\; Liver\;) \end{aligned} \tag{2}$$

This semantic structure lets downstream reasoners interpret clinical context: a clip applier near a major artery triggers safe vascular-management workflows, whereas the same instrument near a nerve bundle raises a mechanical-injury safety alert.

5. The Digital Twin Engine (Layer 3)

The Digital Twin Engine is a volumetric coordinate database maintaining a persistent 3D representation of the operating suite. Whereas Layer 1 works in 2D image coordinates, Layer 3 uses real-time 3D Gaussian Splatting and Signed Distance Fields (SDFs) to reconstruct the cavity's geometry. Every second it updates patient anatomy (surface deformations and retraction into dynamic volumetric meshes), instrument coordinates (3D vectors and jaw angles relative to high-risk structures), active workflow state (blood-loss parameters, tissue states, progression), and critical margins (minimum-distance vectors between energy devices and sensitive boundaries such as the cystic duct or arterial walls). The environment preserves the complete spatial history of the procedure so that no anatomical change or tool manoeuvre is discarded.

6. Multi-Scale Temporal Modelling and Workflow Recognition (Layer 4)

Surgical video exhibits non-uniform temporal dynamics: long repetitive routines punctuated by rapid critical transitions. Standard transformers incur quadratic complexity $O(T^2)$ over long videos, while uniform recurrent models cannot adapt their internal state to varying operational speeds. ORION therefore adopts *SurgicalMamba*, an online phase-recognition model built on the structured state-space duality (SSD) of Mamba-2, with constant linear per-frame complexity $O(d)$ in the model dimension d [6]. Two mechanisms provide long-horizon adaptivity: intensity-modulated temporal stepping and state regramming.

Notation: The symbols used in the formulations of Sections 6 and 13 are defined below and are referenced in that order throughout.

- *t,* τ — wall-clock time and the intrinsic "surgical time" whose speed is warped adaptively.
- $\lambda(t) \in [0,1]$ — continuous intensity (time-warp) signal controlling the rate of state decay.
- *h(t)* — continuous-time hidden state of the state-space model (SSM).
- *A, B(t)* — system decay matrix and input-dependent parameter matrix of the SSM.
- *x(t)* — incoming visual feature vector at time t.
- σ, MLP_λ — sigmoid activation and the compact feed-forward network that infers $\lambda(t)$.
- $\Delta, \alpha_n = 1 + \lambda_n$ — discrete wall-clock step and the corresponding warped-step scaling factor.
- $\bar{A}_n$, $\bar{B}_n$ — discretised state-transition matrices for the warped step.
- $Z^{(c)}$, $S^{(c)}$, I — orthogonal rotation applied at chunk boundary c, its skew-symmetric generator, and the identity matrix.
- $s = [s_1, \ldots, s_k]$, r^*, k — Plackett–Luce score parameters, the ground-truth chronological ordering, and the number of sampled frames.
- I, ϵ, I_t, v_t — latent scene sample, Gaussian noise vector, interpolated latent state, and ground-truth velocity field (SurgWorld).
- u_θ, c, $\mathcal{L}(\theta)$ — velocity network, conditioning semantic prompt, and the flow-matching objective.
- $\hat{a}_t$, $p \in \mathbb{R}^3$, $r \in \mathbb{R}^6$, $g \in [0,1]$ — inferred pseudo-kinematic action, tool-tip Cartesian offset, 6D rotation, and gripper aperture, for left (L) and right (R) manipulators.

6.1 Intensity-Modulated Temporal Stepping

The model distinguishes wall-clock time *t* from an intrinsic surgical time τ, whose speed is warped by a continuous intensity signal $\lambda(t) \in [0,1]$. The continuous-time hidden state *h(t)* evolves as:

$$dh(t)/dt = (1 + \lambda(t))Ah(t) + B(t)x(t) \tag{3}$$

where A is the system decay matrix, $B(t)$ the input-dependent parameter matrix, and $x(t)$ the incoming visual features. The intensity is inferred from local convolutional features by a compact feed-forward network:

$$\lambda(t) = \sigma(\ MLP_{\lambda}(\ x^{conv}\ (t)\ )\ ) \tag{4}$$

with σ the sigmoid. For a discrete step Δ, the transition matrices are adapted to the warped step $\alpha_n \Delta$ (with $\alpha_n = 1 + \lambda_n$):

$$\bar{A}_n = exp(\ \alpha_n \Delta\ A\ ),\ \bar{B}_n \approx \alpha_n \Delta\ B \tag{5}$$

During routine intervals $\lambda(t)$ approaches 0, letting the model slowly decay memory and integrate long-term context; at critical transitions $\lambda(t)$ approaches 1, accelerating decay of past state to prioritise immediate visual feedback.

6.2 State Regramming

To overcome the channel-independence limitation of Mamba-based models, SurgicalMamba applies an input-conditioned orthogonal rotation $Z^{(c)}$ to the accumulated hidden state at each chunk boundary c:

$$h^{(c)} \leftarrow h^{(c)}\ Z^{(c)} \tag{6}$$

This preserves the state norm while re-projecting features into a content-dependent coordinate system. Strict orthogonality is guaranteed by the Cayley transform:

$$Z^{(c)} = (\ I - S^{(c)}\ )(\ I + S^{(c)}\ )^{-1} \tag{7}$$

where $S^{(c)}$ is an input-dependent skew-symmetric matrix ($S^{(c)T} = -S^{(c)}$), forcing the latent space into a block-aligned configuration that mirrors the physical transitions of the surgery.

In implementation, these formulations are integrated into Layer 4 as follows. The continuous-time recurrence is discretised per frame using the warped matrices $\bar{A}_n$ and $\bar{B}_n$, so that each incoming endoscopic frame updates the hidden state in constant time and the model runs online on the streaming GPU pipeline described in Section 14. The intensity network MLPλ is a lightweight convolutional head evaluated alongside the visual encoder, adding negligible latency, while state regramming is applied only at chunk boundaries so that the Cayley-transform inversion is amortised across many frames. The resulting phase posterior is the signal consumed by the predictive engine (Layer 5) and by the Guideline and Risk agents (Layer 8).

6.3 Self-Supervised Pretraining with PL-Stitch

For self-supervised pretraining, ORION uses PL-Stitch, which applies a Plackett-Luce listwise ranking objective to chronological frame sorting [7]. For k randomly sampled frames with score parameters $s = [s_1, \ldots, s_k]$, the probability of recovering the correct ordering r^* is:

$$P(r^* \mid s) = \prod_{i=1}^{k} \frac{\exp(sr_i)}{\sum_{j=i}^{k} \exp(sr_j)} \tag{8}$$

Minimising the negative log-likelihood forces the visual encoder to develop a globally consistent sensitivity to workflow progression and temporal asymmetry.

6.4 Evaluation

Table 4. Performance comparison of surgical workflow recognition models

Metric	LSTM	MS-TCN++	PL-Stitch pretrain	SurgicalMamba
Cholec80 accuracy (%)	88.3	90.1	91.8	94.6
Cholec80 Jaccard (%)	71.4	75.9	78.2	82.7
AutoLaparo accuracy (%)	81.2	84.5	86.9	89.5
AutoLaparo Jaccard (%)	58.1	61.3	64.1	68.9
Processing speed (FPS)	45.0	62.0	95.0	238.7

Table 4 summarises phase-recognition performance on the Cholec80 and AutoLaparo benchmarks. These figures are the results reported in the original publications of the respective methods [6, 7, 8, 10]; they are not the product of a new evaluation carried out in this work, but are reproduced to characterise the temporal component that ORION integrates.

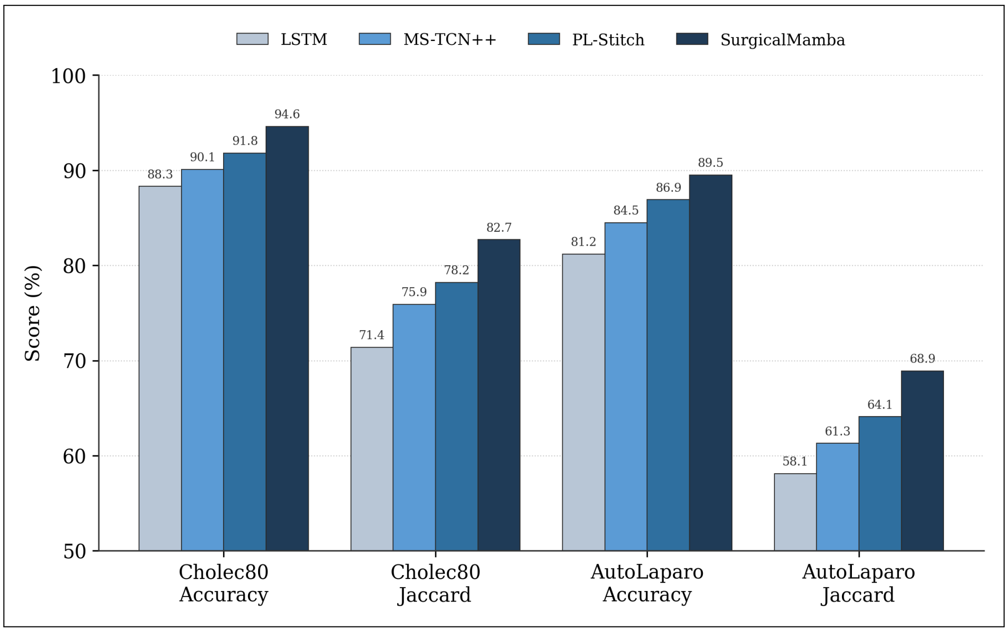

Figure 2. Surgical phase-recognition performance on the Cholec80 and AutoLaparo benchmarks (accuracy and Jaccard) for the LSTM and MS-TCN++ baselines, PL-Stitch pretraining, and SurgicalMamba. Values are reproduced from the methods' original publications

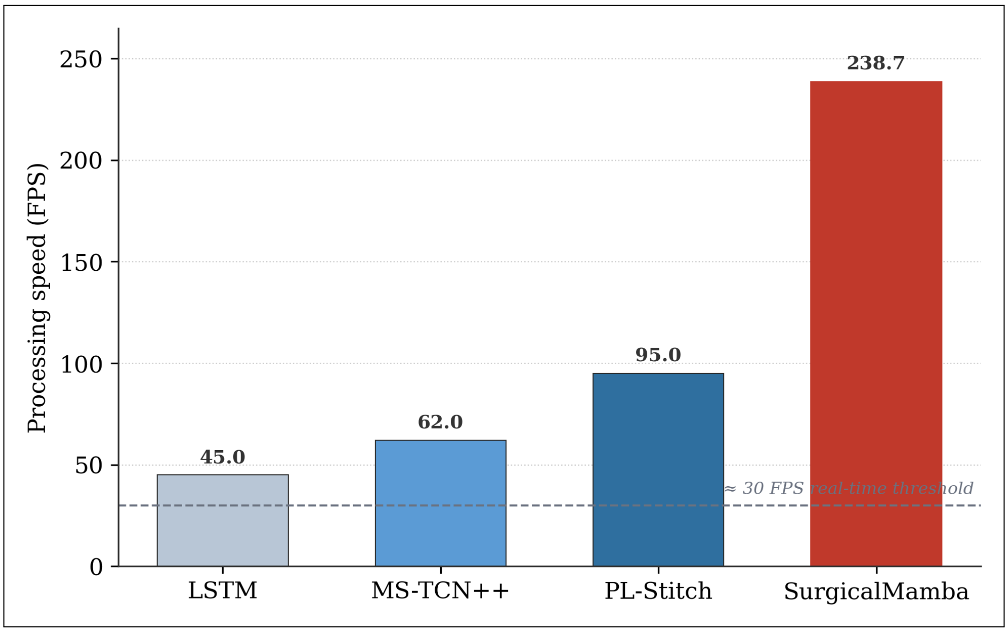

Figure 3. Inference throughput in frames per second for the four temporal models; the dashed line marks the approximate 30 FPS real-time threshold. Values are reproduced from the methods' original publications

ISSN: 2582-337X

7. Predictive Intelligence Engine (Layer 5)

The Predictive Intelligence Engine shifts focus from tracking active events to anticipating upcoming needs and risks. Processing historical phase transitions and real-time state, it estimates remaining procedure duration as a dynamically adjusted distribution ($t_{rem} \pm \sigma$) updated at each transition boundary; a probability of complications from cumulative deviations of tool paths; active bleeding forecasts from blood-loss velocity; instrument-demand profiling to pre-stage scrub-nurse configurations; and operator-fatigue tracking from fine-grained tool tremor and path deviation.

8. Procedural Knowledge Graph (Layer 6)

To provide long-term context, ORION links local procedural events to a multi-institutional clinical knowledge graph implemented in Neo4j and analysed with PyTorch Geometric, structuring experience across millions of historical operations. Entities such as Patient, Disease, Procedure, Complication, Instrument, and Medication are connected by relations (diagnosed_with, underwent, treated_by, resulted_in, utilized, managed_with). The graph continually refines its topology from each completed surgery, letting the platform correlate tool patterns or anatomical anomalies with post-operative outcomes for evidence-based intraoperative support [13, 14].

9. Reinforcement-Learning Support (Layer 7)

The reinforcement-learning support layer does not automate physical maneuvers; instead, it functions as an intraoperative path-planning assistant. Modelling tool-tissue interaction as a Markov Decision Process over expert trajectories from more than 100,000 recorded procedures [17], the policy agent evaluates alternative action paths that minimise tissue strain, maximise visibility, and preserve safety margins, translating kinematic models into real-time guidance on ideal tool placement and path configuration.

10. Multi-Agent Clinical Intelligence and Orchestration (Layer 8)

At the cognitive tier, ORION uses a multi-agent framework orchestrated via LangGraph, modelling clinical analysis as stateful, directed graphs [15, 16]. In place of a single monolithic model, which is susceptible to hallucination, LangGraph coordinates specialised autonomous agents, each optimised for a clinical sub-domain, with state persistence and checkpointing so that latency or disconnection preserves execution state at the last node. Conditional branching adapts the flow to surgical phase, vitals stability, and inter-agent conflict scores.

10.1 Specialist Agent Ecosystem

Table 5. Specialist agents in the ORION multi-agent layer

Agent	Input feeds	Models / paradigm	Clinical output	Safety guardrail
Vision	RGB endoscopic video	Qwen2.5-VL, YOLO12, RT-DETR	Tool detections, masks, tracking vectors	Minimum IoU thresholds; anomaly warnings
Anatomy	Segmented regions, tissue states	nnU-Net, SAM2, MedGemma	Organs, borders, critical margins	Collision warnings near high-risk structures
Instrument	Tool trajectories, gripper states	VideoMAE V2, kinematic monitors	Tracking, usage metrics, state estimates	Velocity filters vs. tracking anomalies
Risk	Scene graphs, bleeding, smoke	PyTorch Geometric hazard models	Bleeding rates, risk trends, alarms	Triple-redundant voting; auditory/visual alerts
Guideline	Phase context, graph state	RAG over SAGES/protocols	Guideline & CVS verification	Compliance checks vs. safety protocols
Literature	PubMed, clinical archives	Dense retrieval, BioBERT	Citations for atypical anatomy	Restricted to peer-reviewed sources
Documentation	Transitions, timelines, usage logs	MedGemma generative text	FHIR/JSON operative reports	Verified vs. EHR registries & logs
Teaching	Frames, graph state, profiles	Explanatory VLMs, GPT-class APIs	Educational overlays for trainees	Separated from primary screen overlays
Analytics	Kinematic paths, durations	Dynamic time warping	Smoothness, motion economy, bottlenecks	Patient-difficulty normalisation

10.2 Inter-Agent Quality Control

A coordinator node continuously computes a cross-agent consensus score. When the Risk Agent detects a deviation from safety margins, or the Guideline Agent flags a non-compliant maneuver, the system raises a priority interrupt and triggers a human-in-the-loop override, requiring explicit confirmation or manual guidance before decision-support resumes.

11. Surgical Memory, Reporting, and Skill Assessment (Layers 9–11)

11.1 Surgical Memory Architecture

Completed operational histories are indexed in a multimodal vector database (Qdrant), converting video sequences, scene graphs, and clinical events into dense embeddings. Clinical teams may subsequently perform similarity searches; for example, "show surgeries where clipping failed" retrieves matching clips, scene graphs, and records, while "find operations similar to this anatomy" matches the active 3D twin mesh against past procedures — turning retrospective archives into searchable clinical resources.

11.2 Automatic Report Generation

Physicians routinely spend twenty minutes or more drafting postoperative notes. ORION compresses this to under thirty seconds: the Documentation Agent synthesises spatial-temporal streams into a JSON-based FHIR payload, normalised against SNOMED CT and ICD-10, covering procedure detail, surgical timeline, active complications, estimated blood loss, instrument-utilisation logs, critical diagnostic findings, and evidence-based suggestions. The payload is written back to the EHR over secure RESTful APIs using OAuth 2.0, perfect forward secrecy, and digital signatures [13].

11.3 Clinical Skill Assessment

ORION provides objective competency tracking from raw video, computing instrument-path smoothness (third-derivative jitter), hand-tremor spectra, a motion-efficiency index, economy of motion, operator idle time, camera-position stability, and cumulative risk-margin deviations. These compile into standardised training scorecards for objective trainee progression [9].

12. Operating-Room Analytics Dashboard (Layer 12)

The top tier aggregates data across suites into an enterprise analytics dashboard, translating clinical streams into operational insight: average phase-specific surgery durations, complication-prevalence trends across teams, instrument-utilisation profiles for tray optimisation, surgeon-efficiency profiling, suite throughput and turnaround, resource-allocation bottlenecks for high-value equipment, and dynamic cost-per-surgery estimation.

13. Generative Surgical World Modelling and Pseudo-Kinematics

A barrier to deploying autonomous decision support is data scarcity: large surgical-video archives lack corresponding instrument kinematics, tool forces, and hand poses. ORION addresses this with a generative physical world model, *SurgWorld*, paired with automated kinematic inference [11].

13.1 SurgWorld Flow-Matching Formulation

SurgWorld is a latent-diffusion model specialised from foundational world models such as Cosmos-Predict [20], fine-tuned with LoRA on the Surgical Action-Text Alignment (SATA) dataset (2,447 expert-annotated clips, > 300,000 frames covering needle grasping, tissue puncture, suture pulling, and knotting). It is trained by flow matching. Given a latent sample I, noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and a normalized time-step $t \in [0,1]$, the interpolated latent is:

$$I_t = (1 - t)\, I + t\, \varepsilon \tag{9}$$

The ground-truth velocity field generating this trajectory is:

$$v_t = \varepsilon - I \tag{10}$$

A network $u_\theta(I_t,\ t,\ c)$ approximates this field, conditioned on an initial frame I_0 and a semantic prompt c, minimising:

$$\mathcal{L}(\theta) = \mathbb{E}_{I,\varepsilon,c,t} \left\| u_\theta(I_t, t, c) - v_t \right\|_2^2 \tag{11}$$

ISSN: 2582-337X

This yields photorealistic, temporally stable, physically plausible video depicting deformation, traction, instrument transitions, and bleeding, generated from text instructions.

Within the framework, this objective is optimised offline. The trained velocity network u_θ is then sampled to synthesise action-labelled clips, which are passed to the inverse-dynamics model of Section 13.2 to recover the pseudo-kinematic vectors that supply the reinforcement-learning support layer (Layer 7) with training trajectories. SurgWorld therefore operates as an offline data-generation component rather than part of the real-time inference path, decoupling the compute-intensive diffusion sampling from the low-latency intraoperative pipeline.

13.2 Inverse Dynamics and Pseudo-Kinematics

To extract policy data from synthesised video, an Inverse Dynamics Model processes temporally adjacent frames (x_t, x_{t+T}) and back-projects observed transitions into continuous pseudo-kinematic actions $\hat{a}_{t:t+T-1}$, represented as a 20-dimensional vector:

$$\hat{a}_t = [\ p_L,\ r_L,\ g_L,\ p_R,\ r_R,\ g_R\ ] \tag{12}$$

where $p \in \mathbb{R}^3$ is the tool-tip Cartesian offset, $r \in \mathbb{R}^6$ is a continuous 6D rotation, and $g \in [0, 1]$ is the gripper aperture for the left (*L*) and right (*R*) manipulators, expressed in the endoscope camera frame.

13.3 Latent Action Quantisation

Where textual descriptions are unavailable, ORION uses Latent Action Quantisation (LAQ) with a Video Joint Embedding Predictive Architecture (V-JEPA) [12]. LAQ encodes discrete visual-transition representations directly from unlabelled archives, yielding a latent-action vocabulary that conditions V-JEPA, enabling self-supervised, action-conditioned pretraining on legacy hospital video without manual annotation.

14. Low-Latency Execution Architecture

To support human-in-the-loop decisions, total round-trip latency from sensor capture through processing to visual or auditory overlay must stay below the ~50 ms human perceptual threshold [3]. ORION decouples streaming sensor execution from discrete batch pipelines using two NVIDIA runtimes: the Holoscan SDK and the Triton Inference Server. Table 6 compares these two runtimes.

The streaming layer divides computation into Operators communicating over explicit ports. A MessageAvailableCondition blocks a downstream operator until a synchronised sensor token arrives, preventing idle cycles, while a DownstreamMessageAffordableCondition halts upstream operators when a downstream queue is full, preventing frame accumulation. Hardware-accelerated memory pipelines pass frames between GPU operators without host-copy, maintaining sub-millisecond inter-operator latency on a single GPU.

Table 6. Holoscan (streaming) vs. Triton (batch) execution runtimes

Dimension	Real-time pipeline (Holoscan)	Multi-model API (Triton)
Primary use	4K streaming, real-time tool tracking, overlays	EHR querying, offline skill assessment, reporting
Input type	Continuous streams (SDI, ST 2110, RTSP)	Discrete batch requests (REST, gRPC, DICOM)
Latency profile	Sub-5 ms inter-operator; sub-30 ms end-to-end	Configurable dynamic batching (5–50 ms)
Memory	Zero-copy pinned GPUDirect allocations	Virtualised host-device copies, tensor queues
Scheduling	Event-driven via strict port conditions	Dynamic batch scheduling, priority queues
Deployment	Edge modules (IGX, Jetson Orin)	Cloud clusters (Kubernetes, AI Enterprise)

15. Regulatory Compliance and Clinical Safety Pathways

Standalone or embedded clinical software is regulated globally as Software as a Medical Device (SaMD) [21]. In India, SaMD falls under the Central Drugs Standard Control Organization (CDSCO) and the Medical Devices Rules (MDR), 2017, which impose a risk-based classification (Classes A–D). Table 7 summarizes this classification hierarchy [22].

Table 7. CDSCO SaMD risk classification (MDR 2017).

Class	**Risk profile**	**Representative software**	**Jurisdiction**	**Primary mandates**
A	Low	Static dashboards, scheduling tools	SLA / import: Central	Literature review, basic QMS
B	Low–moderate	Postoperative skill scoring, timelines	SLA / import: Central	Predicate equivalence, ISO 13485
C	Moderate–high	Real-time phase recognition, auto-reporting	Central Licensing Authority	IEC 62304, ISO 14971, local evaluation
D	High	Active risk prediction, bleeding alerts	Central Licensing Authority	Clinical trials, cybersecurity plan

CDSCO's October 2025 draft guidance classifies standalone software by its potential for patient harm; software that directly guides active interventions (e.g. identifying bleeding rates or forecasting complications in real time) is Class C or D [22]. For these, a Device Master File must document IEC 62304 lifecycle processes, ISO 14971 risk management, an ISO 13485 QMS, in-country clinical performance evaluation across diverse demographics regardless of prior FDA or EU-MDR clearance, and an end-to-end cybersecurity and data-privacy architecture [23, 24, 25].

16. Conclusion

ORION reframes the operating suite from a source of discarded telemetry into a continuously modelled, queryable clinical environment. By layering perception, relational scene understanding, a volumetric digital twin, adaptive temporal modelling, prediction, knowledge grounding, and multi-agent reasoning and by generating its own training signal through SurgWorld when kinematic data is unavailable the framework is designed to close the intraoperative feedback loop while remaining within strict latency and regulatory constraints. As mentioned earlier, realising and validating the integrated system in practice will depend on prospective, in-country clinical evaluation and on the human-in-the-loop safeguards that keep automated inference advisory rather than autonomous.

References

[1] Özsoy, Ege, Chantal Pellegrini, Tobias Czempiel, Felix Tristram, Kun Yuan, David Bani-Harouni, Ulrich Eck, Benjamin Busam, Matthias Keicher, and Nassir Navab. ”Mm-or: A Large Multimodal Operating Room Dataset for Semantic Understanding of High-Intensity Surgical Environments.” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025, 19378-19389.

[2] Özsoy, Ege, Arda Mamur, Felix Tristram, Chantal Pellegrini, Magdalena Wysocki, Benjamin Busam, and Nassir Navab. ”Egoexor: An Ego-Exo-Centric Operating Room Dataset for Surgical Activity Understanding.” Advances in Neural Information Processing Systems 2026, vol. 38.

[3] NVIDIA. (2026). Deploy NVIDIA Holoscan: Real-Time Sensor AI for Medical Imaging and Industrial Inspection. NVIDIA Developer. https://www.spheron.network/blog/nvidia-holoscan-gpu-cloud-sensor-ai-guide/

[4] Shin, Jongmin, Enki Cho, Ka Young Kim, Jung Yong Kim, Seong Tae Kim, and Namkee Oh. ”Towards Holistic Surgical Scene Graph.” In International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham: Springer Nature Switzerland, 2025, 617-626.

[5] Henriques, Angelo, Korab Hoxha, Daniel Zapp, Peter C. Issa, Nassir Navab, and M. Ali Nasseri. ”Decoding the Surgical Scene: A Scoping Review of Scene Graphs in Surgery.” Medical Image Analysis 2026, 104083.

[6] Oh, Sukju, and Sukkyu Sun. ”SurgicalMamba: Dual-Path SSD with State Regramming for Online Surgical Phase Recognition.”, 2026, arXiv preprint arXiv:2605.14889.

[7] Che, Chengan, Chao Wang, Xinyue Chen, Sophia Tsoka, and Luis C. Garcia-Peraza-Herrera. ”A Stitch in Time: Learning Procedural Workflow Via Self-Supervised Plackett-Luce Ranking.” IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026, 17000-17010.

[8] Wang, Ziyi, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and

Yunhui Liu. ”Autolaparo: A New Dataset of Integrated Multi-Tasks for Image-Guided Surgical Automation in Laparoscopic Hysterectomy.” In International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham: Springer Nature Switzerland 2022, 486-496.

[9] Rueckert, Tobias, David Rauber, Raphaela Maerkl, Leonard Klausmann, Suemeyye R. Yildiran, Max Gutbrod, Danilo Weber Nunes et al. ”Comparative Validation of Surgical Phase Recognition, Instrument Keypoint Estimation, and Instrument Instance Segmentation in Endoscopy: Results of the PhaKIR 2024 challenge.” Medical Image Analysis 2026, 103945.

[10] Long, Jun, Junkun Hong, Zidong Wang, Tingxuan Chen, Yunfei Chen, and Liu Yang. ”SPHASE: Multi-Modal and Multi-Branch Surgical Phase Segmentation Framework Based on Temporal Convolutional Network.” IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2023, 586-593.

[11] He, Yufan, Pengfei Guo, Mengya Xu, Zhaoshuo Li, Andriy Myronenko, Dillan Imans, Bingjie Liu et al. ”Cosmos-H-Surgical: Learning Surgical Robot Policies from Videos via World Modeling.”, 2025, arXiv preprint arXiv:2512.23162.

[12] Scalable Surgical World Modeling via Latent Action Prediction, 2025. OpenReview - https://openreview.net/forum?id=aYQYOVm2AB

[13] Lee, Gyubok, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, and Jong Ha Lee. ”FHIR-Agentbench: Benchmarking LLM Agents for Realistic Interoperable EHR Question Answering.”, 2025, arXiv preprint arXiv:2509.19319.

[14] De Maio, Carmen, Giuseppe Fenza, Domenico Furno, Teodoro Grauso, and Vincenzo Loia. ”Privacy-Preserving Healthcare Data Interactions: A Multi-Agent Approach Using LLMs.” Journal of Communications Software and Systems 2025, vol. 21, no. 1, 13-22.

[15] Borkowski, Andrew A., and Alon Ben-Ari. ”Multiagent AI Systems in Health Care: Envisioning Next-Generation Intelligence.” Federal Practitioner 2025, vol. 42, no. 5, 188.

[16] Agent orchestration: LangGraph vs CrewAI vs Temporal. (2026). Nirmitee.io Technical Report - https://nirmitee.io/blog/langgraph-crewai-temporal-custom-orchestration-

healthcare-agents-2026/

[17] Fazzari, Edoardo, and Cesare Stefanini. ”Deep Reinforcement Learning for Surgical Robotics with State and Image Information: A Survey.” (2026).

[18] Ravi, Nikhila, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr et al. ”Sam 2: Segment Anything in Images and Videos.” In International Conference on Learning Representations 2025, vol. 2025, 28085-28128.

[19] Isensee, Fabian, Paul F. Jäger, Simon AA Kohl, Jens Petersen, and Klaus H. Maier-Hein. ”Automated Design of Deep Learning Methods for Biomedical Image Segmentation.”, 2019 arXiv preprint arXiv:1904.08128.

[20] NVIDIA. (2024). Cosmos World Foundation Model Platform for Physical AI. NVIDIA. https://research.nvidia.com/publication/2025-01_cosmos-world-foundation-model-platform-physical-ai

[21] Central Drugs Standard Control Organization. (2017). Medical Devices Rules, 2017. Government of India. https://cdsco.gov.in/opencms/resources/UploadCDSCOWeb/2022/m_device/Medical%20Devices%20Rules,%202017.pdf

[22] Central Drugs Standard Control Organization. (2025). Draft Guidance Document On Medical Device Software Under MDR 2017. Government of India. https://cdsco.gov.in/opencms/export/sites/CDSCO_WEB/Pdf-documents/Guidance-document-on-Medical-Device-Software-under-MDR-2017.pdf

[23] International Electrotechnical Commission. (2015). IEC 62304: Medical device software - Software life-cycle processes. IEC.

[24] International Organization for Standardization. (2019). ISO 14971: Medical devices - Application of risk management to medical devices. ISO.

[25] International Organization for Standardization. (2016). ISO 13485: Medical devices - Quality management systems - Requirements for regulatory purposes. ISO.