MedOS: AI-XR-Cobot World Model for Clinical Perception and Action

Yingcheng Charles Wu[1], Ming Yin[2], Baiyu Shi[3], Zaixi Zhang[1], Di Yin[1], Xiaotong Wang[1], Youjuan Wang[1], Jigang Fan[1], Ruofan Jin[1], Hanchen Wang[4], Kejun Albert Ying[5], Kuan Pang[4], Rebecca Rojansky[1], Christina Curtis[1,6], Zhenan Bao[3], Mengdi Wang[2*], Le Cong[1*]

[1] Department of Pathology, Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[2] Princeton AI Lab, Department of Electrical & Computer Engineering, Princeton University, Princeton, NJ, USA

[3] Department of Chemical Engineering, Stanford University, Stanford, CA, 94305, USA

[4] Department of Computer Science, Stanford University, Palo Alto, CA, USA

[5] The Phil and Penny Knight Initiative for Brain Resilience, Stanford University; Institute for Protein Design, University of Washington, Seattle, WA, USA

[6] Department of Medicine, Department of Biomedical Data Science, Stanford Cancer Institute, Stanford School of Medicine, Stanford, CA, USA.

Corresponding:
Le Cong (congle@stanford.edu), Mengdi Wang (mengdiw@princeton.edu)

Running title: MedOS: AI-XR-Cobot World Model

Abbreviations:

AI: Artificial Intelligence; XR: Extended Reality; LLM: Large Language Model; VLM: Vision-Language Model; OR: Operating Room; EHR: Electronic Health Records; GRPO: Group Relative Policy Optimization; MSV: MedSuperVision.

Abstract

Medicine historically separates abstract clinical reasoning from physical intervention. We bridge this divide with MedOS, a general-purpose embodied world model. Mimicking human cognition via a dual-system architecture, MedOS demonstrates superior reasoning on biomedical benchmarks and autonomously executes complex clinical research. To extend this intelligence physically, the system simulates medical procedures as a physics-aware model to foresee adverse events. Generating and validating on the MedSuperVision benchmark, MedOS exhibits spatial intelligence for reasoning and action. Crucially, we demonstrate that this platform democratizes clinical expertise and narrows the performance gap between junior and senior physicians. MedOS transforms clinical intervention towards a collaborative discipline where human intuition and machine intelligence co-evolve.

Keywords: World Models, Spatial Intelligence, Embodied AI, Surgical Robotics, Extended Reality (XR)

INTRODUCTION

Medicine relies on the integration of clinical reasoning to diagnose disease and physical execution to intervene. While recent advances in artificial intelligence have transformed the reasoning domain, with large language models potentially achieving expert proficiency in medical licensing exams and diagnostic dialogue, the physical domain of interventions remains a critical bottleneck[1,2]. Clinical outcomes depend not merely on static knowledge but on perception, dexterity, and real-time decision-making under uncertainty. Current medical AI remains largely disembodied and confined to the digital analysis of electronic health records or static imaging, leaving it unable to perceive or act in the dynamic reality of procedural medicine [3,4]. Conversely, surgical robotics provide precision tele-operation but potentially remain unintelligent that are blind to anatomical context and somewhat dependent on human control [5,6].

To bridge this fundamental divide, we introduce MedOS, a unified collaborative intelligence platform that renders clinical environments perceivable and operable by AI. MedOS represents a conceptual shift from passive data analysis to the idea of embodied world model. It integrates agentic reasoning with extended reality enabled multimodal interfaces and robotic control systems to create an end-to-end framework that links longitudinal patient history in the digital world to real-time interaction in the physical world. By grounding abstract medical knowledge into a dynamic state space, MedOS enables the AI to function not merely as a consultant but as a perceiving co-physician capable of active collaboration.

The architectural innovation of MedOS lies in its mimicry of expert human cognition through a dual-system mechanism [7,8]. In clinical practice, a physician or surgeon should seamlessly toggle between expert strategy and deliberate action. MedOS operationalizes this by employing a system 2 slow agent to process macro-context such as demographics and meso-context such as perioperative plans, while simultaneously deploying a system 1 fast agent to handle millisecond-level risk perception and reflex-like guidance. This architecture allows the AI to simulate a physics model by reasoning about force vectors, predicting tissue responses, and identifying adverse events such as bleeding risks in real-time.

In this work, we present an end-to-end instantiation of MedOS for the interventional domain. On the reasoning front, the system achieves state-of-the-art accuracy on challenging biomedical benchmarks and outperforms frontier models through a self-evolving critique loop. To enable physical perception, we constructed MedSuperVision, a large-scale benchmark of egocentric surgical videos annotated with expert narratives and instrument dynamics. Recognizing that general-purpose vision-language models struggle with the subtle textures and depth of biological tissue, we trained a domain-specialized world model using group relative policy optimization. This training enables MedOS to decode visual input from extended reality glasses to execute counterfactual prediction, foreseeing potential margin violations or tissue tears before they materialize. Furthermore, we demonstrate that this spatial intelligence can be directly translated into action, empowering autonomous robotic systems with stability and enabling real-time XR-human-robot collaboration that significantly enhances surgical efficiency. By endowing AI with the ability to think with clinical rigor and see with surgical precision, MedOS advances the field toward autonomous and reproducible healthcare where human intuition and machine intelligence co-evolve to assist patient care.

RESULTS

MedOS: An Agentic World Model across Digital and Physical Scales

Current medical systems largely work in isolation, lacking the capability to unify abstract clinical reasoning with physical intervention [8,9]. To bridge this gap, MedOS creates a unified agentic world model that connects the digital world of longitudinal history with the physical world of surgery (Figure 1A). The architecture functions across two fundamental planes, Digital and Physical, to align medical logic with procedural reality.

At Level 1 (Digital World), the system operates within a semantic and clinical logic framework to establish the strategic baseline. It integrates Step 1 (Macro-Context) to process lifelong tokens, identifying patient phenotypes such as cirrhosis and risks like portal hypertension. This converges with Step 2 (Meso-Context) for perioperative planning, where MedOS analyzes recent clinical events to detect states like coagulopathy and unstable hemodynamics, explicitly formulating a plan to minimize tissue trauma and ensure strict hemostasis.

Crucially, the architecture transitions into Level 2 (Physical World), a domain of embodied and spatial intelligence designed for XR-Robot-Human Collaboration. Through high-bandwidth XR Streaming and Robotic Control interfaces (controlling instruments like a laparoscope), MedOS models a 3D state space that includes the egocentric view, real-time scene depth, and instrument interactions. To master the high-stakes dynamics of Step 3 (Micro-Execution), we implemented a dual-system architecture: First, a reflex-like System 1 (Fast) module processes real-time streams to detect immediate adverse events. For instance, upon perceiving fibrotic adhesions, it reasons that the tissue is friable with high tear risk with traction, and immediately guides the robotic action to use suction dissection (avoid grasper). Second, A deliberative System 2 (Slow) module coordinates high-level planning and trajectory optimization based on the full digital context. Such architecture helps the XR-Cobot-Human collaboration in a unified space.

MedOS builds on a specialized multi-agent framework. A Coordinator Agent orchestrates the workflow by decomposing complex queries for specialized modules, including EHR, Guideline, Radiology, and Pathology agents. The core Reasoning Agent executes a structured thinking template driven by evidence synthesis and causal inference. This process is governed by a Self-evolving Critic Agent that continuously evaluates plans, supported by a tool ocean providing capabilities like EHR reasoning or clinical research planning.

We validated this architecture on challenging biomedical benchmarks, where MedOS consistently establishes a new state-of-the-art. On MedQA (USMLE) [10], MedOS achieves an accuracy of approximately 97%, surpassing frontier models such as Gemini 3 Pro (~95%) and GPT-5.2 Thinking (~96%) (Figure 1B). Similarly, on the GPQA benchmark for expert-level reasoning [11], MedOS scores ~94%, maintaining a robust lead over Claude 4.5 Opus (~90%) (Figure 1C). Furthermore, MedOS demonstrated inference-time scaling properties (Figure 1D); by increasing the Token budget for the system 2 thinking process (from 1x to 9x), the model's performance systematically improves, providing direct evidence that the dual-system design enables the AI to evolve its strategies for complex clinical scenarios.

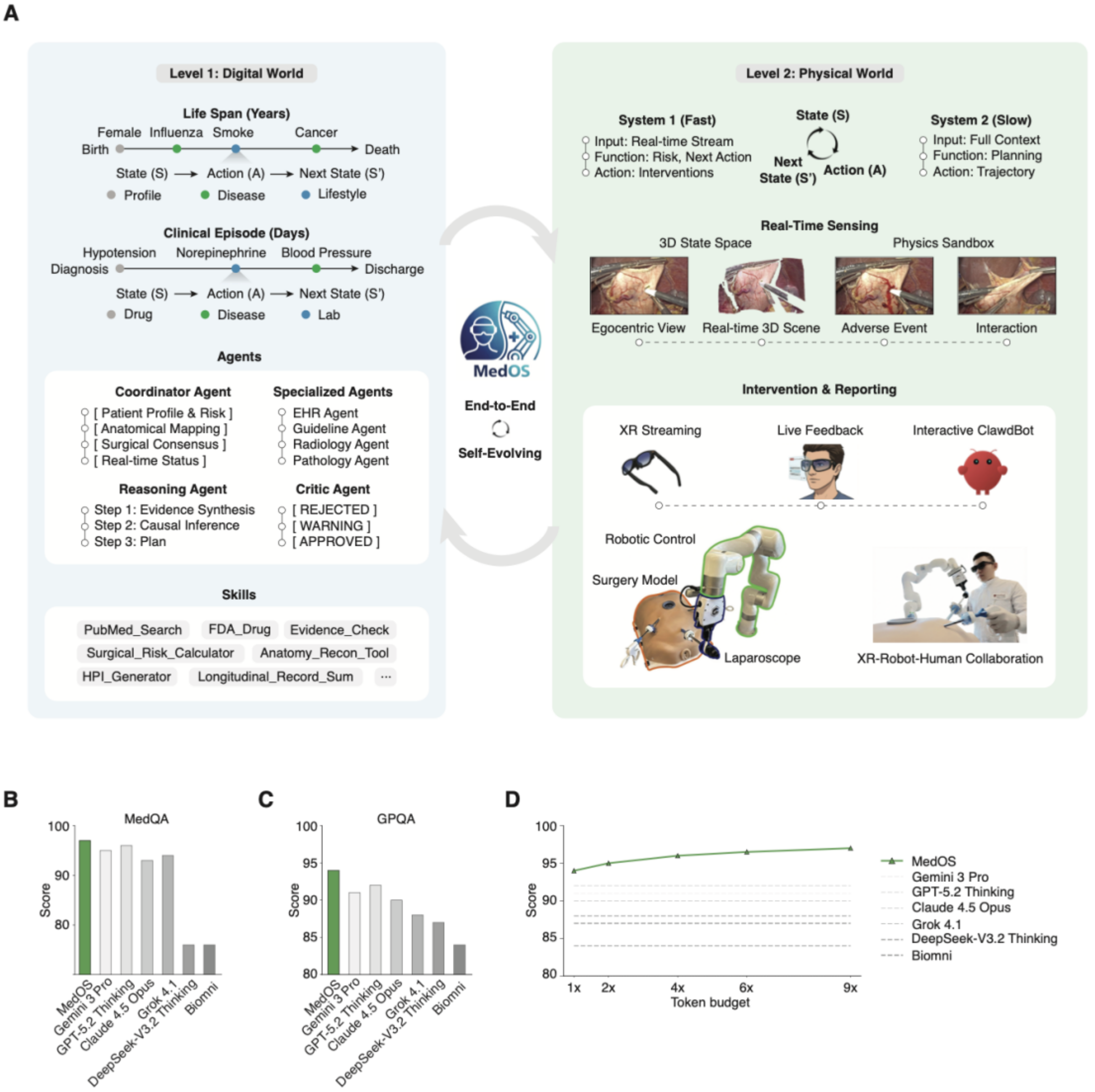


Figure 1. MedOS: An Agentic World Model across Digital and Physical Scales. (A) Schematic overview of the MedOS architecture. The framework bridges two fundamental planes: Level 1 (Digital World) and Level 2 (Physical World). Level 1 integrates semantic and clinical logic, processing longitudinal patient history (Macro-Context) and perioperative planning (Meso-Context) to form strategic plans. Level 2 represents the domain of embodied and spatial intelligence, designed to

facilitate XR-Robot-Human Collaboration. It connects high-bandwidth XR streaming with Robotic Control (e.g., Laparoscope) to simulate a real-time physics model. This level utilizes a dual-system cognitive architecture: a System 1 (Fast) agent for reflexive risk perception and immediate intervention, and a System 2 (Slow) agent for high-level trajectory planning. The bottom panel illustrates the multi-agent workflow, orchestrated by a Coordinator Agent and optimized by a self-evolving Critic Agent. (B and C) Comparative evaluation of reasoning capabilities. MedOS achieves state-of-the-art accuracy on the (B) MedQA (USMLE) and (C) GPQA benchmarks, significantly outperforming frontier models including Gemini 3 Pro and GPT-5.2 Thinking. (D) Inference-time scaling analysis. The graph demonstrates the positive correlation between the token budget allocated for system 2 thinking processes and model performance, validating the efficacy of the deliberative reasoning module. Models (e.g., Gemini 3 Pro, Claude 4.5 Opus) which are evaluated in a fixed zero-shot setting (dashed lines), MedOS (green line) dynamically utilizes increased test-time compute to refine its reasoning path. All photographs shown are of the authors.

MedOS Reasons Across Diverse Tasks and Democratizes Clinical Expertise

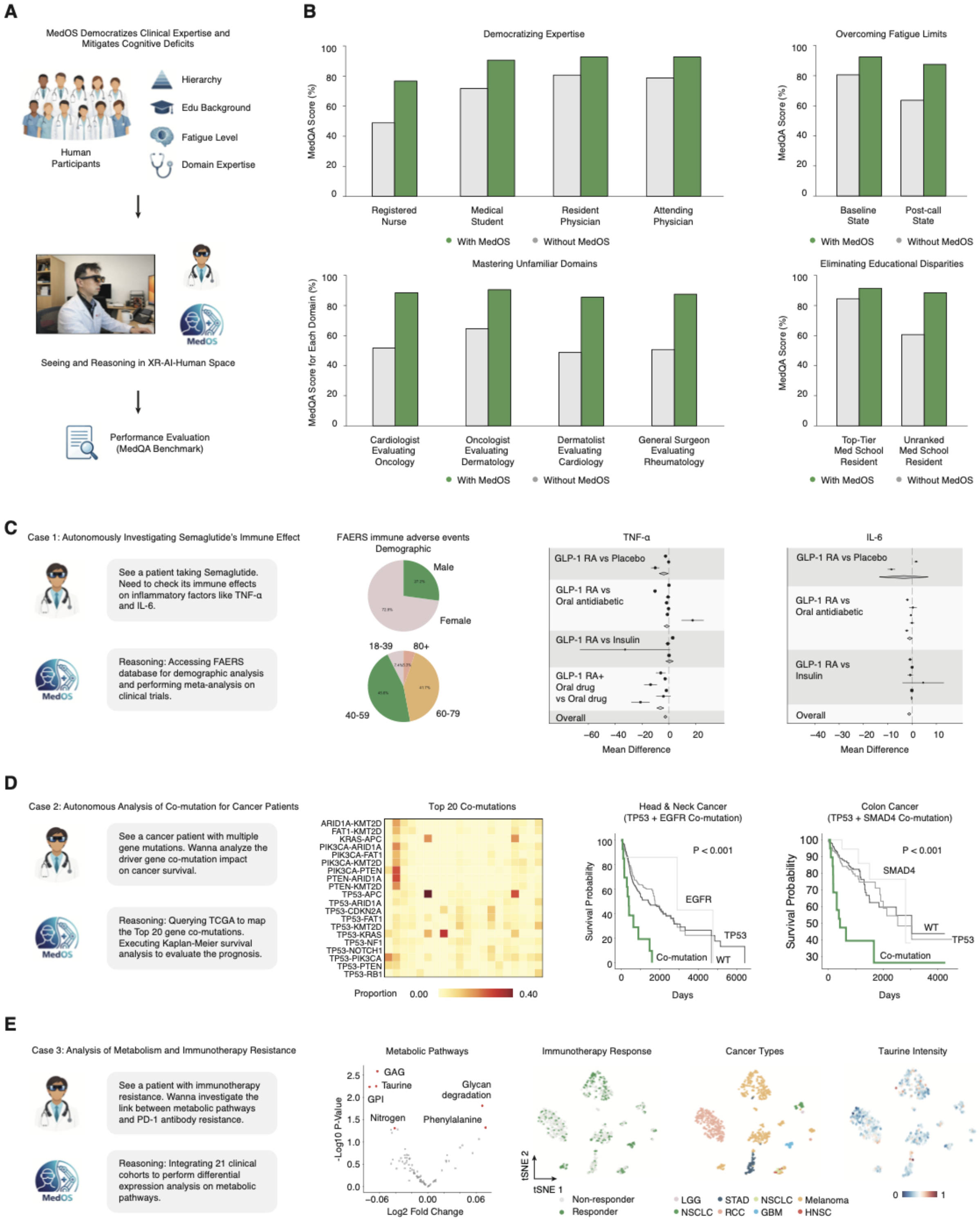
To evaluate the capacity of our system to augment human intelligence and mitigate inherent cognitive limitations, we designed a human-AI collaboration study involving participants (n = 24) with varying levels of hierarchy, educational backgrounds, and physiological states (Figure 2A). The quantitative results indicate a profound leveling effect across all dimensions of expertise (Figure 2B). We observed that MedOS enabled registered nurses to improve their diagnostic accuracy from an unaided score of 49% to an AI-augmented score of 77%, while medical students advanced from 72% to 91%. Remarkably, this augmentation allowed less experienced cohorts to rival or surpass the performance of attending physicians, who scored 79% unaided and reached a ceiling of 93% with assistance. Similarly, resident physicians improved from a baseline of 81% to 93%, suggesting that the system effectively bridges the knowledge gap inherent in traditional medical training hierarchies.

Beyond baseline expertise, we investigated the ability of the system to counteract cognitive deficits caused by fatigue and educational disparities. While sleep-deprived post-call physicians suffered a significant drop in performance to 64% compared to their baseline state of 81%, the integration of MedOS restored their accuracy to 88%, surpassing even their well-rested baseline. Furthermore, the model effectively closed educational gaps; the lower-performance cohort from unranked schools saw a dramatic rise from 61% to 89%, narrowing the difference with the high-performance cohort that improved from 85% to 92%. The model also mastered unfamiliar domains, enabling specialists to operate with high precision outside their core disciplines. For instance, cardiologists evaluating oncology cases improved from 52% to 89%, and oncologists assessing dermatology questions rose from 65% to 91%. Even in highly distinct pairings, such as dermatologists evaluating cardiology or general surgeons assessing rheumatology, accuracy surged from 49% to 86% and 51% to 88% respectively. These data demonstrate the capability of MedOS to democratizes clinical expertise across doctor with varying levels of hierarchy, educational backgrounds, and physiological states as well as across medical students and nurses.

We next extended the capability of MedOS from passive question answering to autonomous clinical research, tasking the model with executing complex workflows that span from user queries to data-driven report generation (Figure 2C-E). In the first case, a human user requested an investigation into the immune side effects of Semaglutide. In response, the MedOS coordinator agent formulated a multi-step research plan and deployed specialized tools to access the FDA Adverse Event Reporting System (FAERS) database[12] for a demographic analysis of immune adverse events. The system then autonomously executed a meta-analysis on clinical trial data, generating forest plots that revealed a statistically significant reduction in TNF-alpha and IL-6 levels in the GLP-1 RA group compared to placebo and insulin controls. This process demonstrates the ability of the model to synthesize raw pharmacovigilance data into actionable clinical evidence (Figure 2C).

We next applied the system to genomic oncology to assess the prognostic implications of driver gene co-mutations (Figure 2D). The workflow commenced with a user request to analyze a cancer patient presenting with multiple gene mutations and to evaluate the impact of co-mutations on survival outcomes. MedOS responded by querying The Cancer Genome Atlas (TCGA)[13] to map the top 20 gene co-mutations, visualizing the frequency of interactions such as TP53-APC and KRAS-APC via a heatmap. Proceeding to survival analysis, the agent utilized statistical tools to generate Kaplan-Meier curves, which uncovered that patients with TP53 and EGFR co-mutations in head and neck cancer, as well as those with TP53 and SMAD4 co-mutations in colon cancer, faced significantly worse survival probabilities ($P < 0.001$) compared to single-mutation or wild-type cohorts, partly consistent with prior studies[14,15]. These findings highlight the capacity of the system to perform complex bioinformatics tasks and stratify patient risk profiles based on high-dimensional genomic data.

Finally, we explored the complex mechanisms of immunotherapy resistance by linking metabolic pathways to PD-1 antibody efficacy (Figure 2E). The session began with a clinician inquiring about the connection between tumor metabolism and immune checkpoint inhibitor resistance. In response, MedOS integrated data from 21 clinical cohorts[16] to perform a differential expression analysis, successfully identifying metabolic pathways such as glycosaminoglycan (GAG) biosynthesis and taurine metabolism as significantly enriched in non-responders. Interestingly, the system generated t-SNE projections visualizing the single-cell clustering of responders versus non-responders, explicitly mapping the intensity of taurine metabolism to the resistant phenotype across melanoma and NSCLC samples. This case illustrates the potential of MedOS to function as a co-investigator that can uncover novel biological mechanisms and suggest therapeutic targets by autonomously integrating multi-omics data.



Case 1:

See a patient taking Semaglutide. Need to check its immune effects on inflammatory factors like TNF-α and IL-6.

Reasoning: Accessing FAERS database for demographic analysis and performing meta-analysis on clinical trials.

Case 2:

See a cancer patient with multiple gene mutations. Wanna analyze the driver gene co-mutation impact on cancer survival.

Reasoning: Querying TCGA to map the Top 20 gene co-mutations. Executing Kaplan-Meier survival analysis to evaluate the prognosis.

Case 3:

See a patient with immunotherapy resistance. Wanna investigate the link between metabolic pathways and PD-1 antibody resistance.

Reasoning: Integrating 21 clinical cohorts to perform differential expression analysis on metabolic pathways.

Figure 2. MedOS Reasons Across Diverse Medical Tasks and Democratizes Clinical Expertise. (A) Schematic of the human-AI collaboration study. Participants with varying hierarchies, educational backgrounds, and fatigue levels utilize MedOS in an XR-AI-Human space to solve clinical problems (MedQA), aiming to evaluate the system's ability to mitigate cognitive deficits. (B) Quantitative evaluation on the MedQA benchmark across four critical dimensions. The bar charts demonstrate that MedOS: (1) Democratizes Expertise, elevating Registered Nurses and Medical Students to performance levels comparable to Attending Physicians; (2) Overcomes Fatigue Limits, significantly restoring the performance of sleep-deprived (Post-call) physicians; (3) Masters Unfamiliar Domains, enabling specialists (e.g., Cardiologists) to achieve high accuracy in out-of-distribution fields (e.g., Oncology); and (4) Balances Educational Disparities, closing the gap between residents from top-tier and unranked medical schools. (C–E) Demonstration of Autonomous Clinical Research capabilities across diverse tasks. (C) Case 1: Investigation of Semaglutide's immune effects. MedOS autonomously accesses the FAERS database for demographic analysis and performs meta-analysis on inflammatory factors (TNF-α and IL-6). (D) Case 2: Analysis of cancer co-mutations. The system queries TCGA data to map top gene co-mutations and executes Kaplan-Meier survival analysis for Head & Neck and Colon cancers. (E) Case 3: Exploration of immunotherapy resistance. MedOS integrates clinical cohorts to perform differential expression analysis on metabolic pathways, visualizing the link between metabolites (e.g., Taurine) and PD-1 antibody response via t-SNE projections. All photographs shown are of the authors.

Training MedOS To See and Reason with Spatial Intelligence

To enable the AI to perceive the physical reality of surgery beyond static frames, we manually assembled MedSuperVision (MSV), a large-scale, expert-annotated surgical video dataset from open-access educational resources, designed to benchmark clinical operation understanding (Figure 3A). We devised a rigorous four-phase curation protocol: Phase I aggregated diverse egocentric videos enriched by surgeon's narrative and expert commentary to capture intent; phase II processed this raw footage through time-frame segmentation, chain-of-thought extraction, and patient health information removal, resulting in a split of 80% for training and 20% for validation. After quality control, the dataset contains videos of 85,398 minutes, spans multiple disciplines, dominated by hepatobiliary and gastrointestinal surgeries but also covering urologic, vascular, and thoracic procedures (Figure 3B). We observed a realistic video duration distribution, ranging from short clips (<10 min) to extended procedures (>120 min), with the majority falling in the 60-120 minute range. All these videos feature narrations by 1,882 clinical experts describing each surgery.

Following the dataset construction, we utilized this benchmark to train the MedOS world model using a dual-system training strategy (Figure 3A, Phase III). We employed supervised fine-tuning (SFT) followed by group relative policy optimization (GRPO)[17] based on Qwen3-VL-8B-Instruct[18] to distinctively optimize two sub-modules: a system 1 (fast) module trained for immediate next action and risk detection, and a system 2 (slow) module optimized for trajectory & chain-of-thought planning. We observed through comparative evaluation that while frontier models like Gemini 3 Pro falter in dynamic tasks, MedOS outperforms baselines by consistent margins. For example, in system 1 benchmarks (Figure 3C), MedOS achieves balanced instrument recall and leads significantly in contact detection rate (~85% vs. ~80%) and action recognition rate (~82% vs. ~75%).

We further noted that the performance gap widens significantly when evaluating complex reasoning on Unseen Disciplines (Figure 3E). In the radar chart analysis, MedOS demonstrates superior capability in risk reasoning and next step prediction, scoring between ~80-90%, whereas the general Gemini 3 Pro model drops to ~70%. MedOS maintains a distinct advantage in causal inference and context reasoning. Finally, to validate clinical utility, we conducted a blinded human expert rating (Figure 3F), composed of 5 licensed medical doctors who rated 100 scenes for MedOS and Gemini 3 Pro. As a result, MedOS was declared the winner in approximately 60% of test cases, compared to a ~15% win rate for Gemini 3 Pro and a ~25% tie rate. Together, these data validated MedOS' application in real-world surgical interpretation.

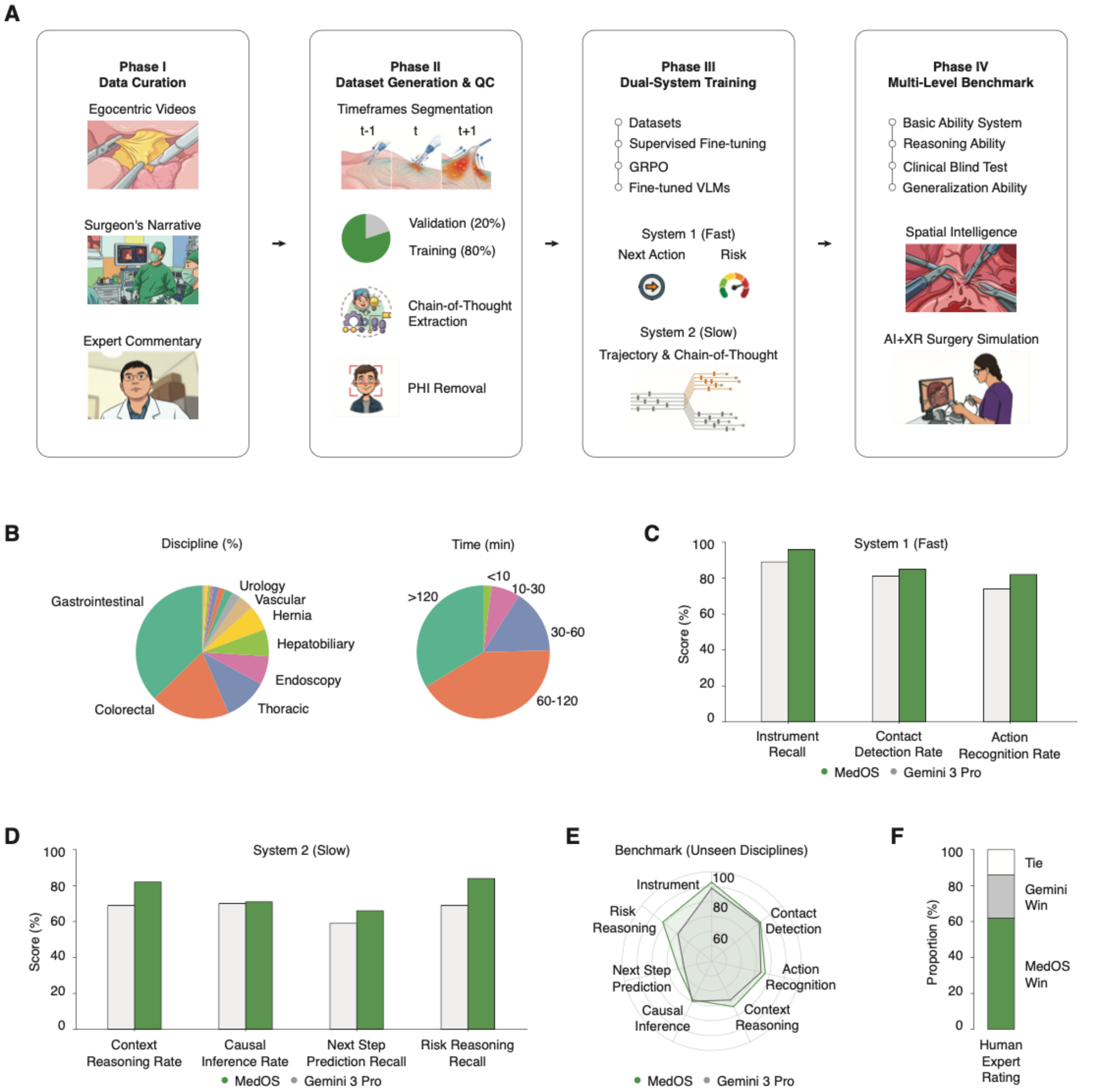


Figure 3. MedSuperVision: A Large-Scale Benchmark and Dual-System Training for Physical Perception (A) The MedSuperVision (MSV) data curation and training pipeline. The protocol involves four phases: (I) aggregation of egocentric videos with expert narratives; (II) dataset generation with time-frame segmentation and chain-of-thought extraction; (III) dual-system training using Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO) to distinctively optimize system 1 (risk/action) and system 2 (trajectory/reasoning) modules; and (IV) multi-level benchmarking. (B) Distribution of the MSV dataset by surgical discipline (left) and procedure duration (right), highlighting diversity across Hepatobiliary, Gastrointestinal, and Urology specialties. (C) Quantitative assessment of system 1 (Fast) capabilities. MedOS outperforms the baseline (Gemini 3 Pro) in high-frequency

tasks, including instrument recall, contact detection, and action recognition. (D) Quantitative assessment of system 2 (Slow) capabilities. MedOS demonstrates superior performance in complex cognitive tasks such as context reasoning and causal inference. (E) Radar chart illustrating model generalization on unseen surgical disciplines. MedOS maintains robust performance across all axes compared to the baseline. (F) Clinical validation via blinded human expert rating. MedOS-generated guidance was preferred in approximately 60% of test cases.

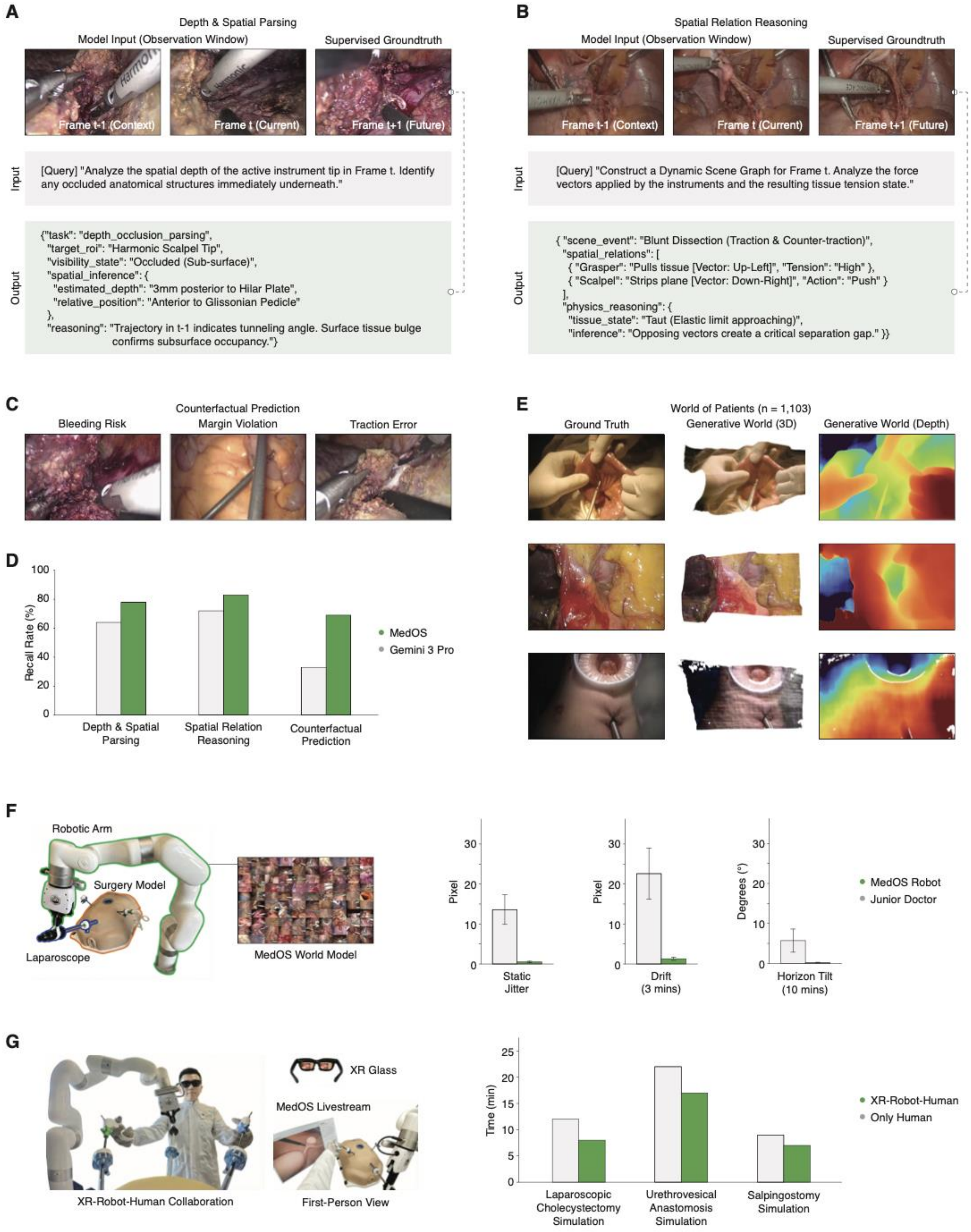
MedOS Unlocks Spatial Intelligence for Physics-Aware Surgical Reasoning

To decode the latent 3D structure of the operating field from 2D egocentric frames, we positioned MedOS to demonstrate Spatial Intelligence, the ability to not just recognize objects, but to understand their 3D position, mechanical interactions, and causal consequences. In depth & spatial parsing tasks (Figure 4A), the model inputs a temporal observation window to resolve spatial ambiguities. We observed that it successfully parses occlusion states, estimating the harmonic scalpel tip depth posterior to hilar plate and localizing it relative to hidden critical structures like the glissonian pedicle. Notably, the model grounds this estimation in physical cues, reasoning that the trajectory in indicates tunneling angle and surface tissue bulge confirms subsurface occupancy.

We next analyzed the forces and mechanics of surgery via dynamic scene graphs to understand how physical actions alter the anatomical environment (Figure 4B). We inferred spatial relation reasoning by categorizing complex maneuvers such as blunt dissection involving traction and counter-traction. Specifically, the model decomposes this action into precise force vectors, identifying that the grasper pulls tissue while the scalpel strips the plane. Crucially, it evaluates the resulting tissue state as taut or approaching its elastic limit, deriving the physics-based inference that opposing vectors create a critical separation gap. This ability to translate visual data into mechanical forces demonstrates that MedOS partly understands the physical consequences of instrument interaction beyond mere semantic classification.

We finally examined MedOS as a predictive world model capable of counterfactual prediction, simulating hypothetical scenarios to foresee adverse events before they materialize (Figure 4C). We observed that it successfully predicts potential failures such as bleeding risk, margin violation, and traction error based on current instrument trajectories. Quantitative evaluation validated these capabilities (Figure 4D), where MedOS achieves recall rates of approximately 82% in spatial relation reasoning and 78% in depth and spatial parsing, significantly outperforming the baseline Gemini 3 Pro which scores between 60% and 70%. The performance gap is most pronounced in counterfactual prediction, where MedOS achieves a recall of 68% compared to the baseline 32%, highlighting the specialized world model superior ability to anticipate physical consequences in high-stakes environments. These quantitative results establish the superiority of a specialized world model in anticipating physical consequences within high-stakes environments.

To further validate the fidelity of our world model, we deployed MedOS to perform generative world reconstruction across a large-scale cohort of 1,103 patients (Figure 4E). By synthesizing high-fidelity 3D representations and dense depth maps from sparse egocentric video inputs, the system reconstructed the complex topography of surgical fields ranging from uterine cavities to vascular beds. This large-scale generation provides a crucial digital twin of diverse patients, serving as a high-fidelity environment where robotic agents can be trained and tested without risk to human patients. The comparison between ground truth and generated worlds confirms that the model captures fine-grained textural details and geometric depth essential for realistic simulation. Ultimately, this capacity for large-scale 3D reconstruction transforms the system from a passive analyzer into a generative engine capable of creating immersive training environments for both human surgeons and autonomous robots.



[Query] "Analyze the spatial depth of the active instrument tip in Frame t. Identify any occluded anatomical structures immediately underneath."

```
{"task": "depth_occlusion_parsing",
  "target_roi": "Harmonic Scalpel Tip",
  "visibility_state": "Occluded (Sub-surface)",
  "spatial_inference": {
    "estimated_depth": "3mm posterior to Hilar Plate",
    "relative_position": "Anterior to Glissonian Pedicle"
  },
  "reasoning": "Trajectory in t-1 indicates tunneling angle. Surface tissue bulge
                 confirms subsurface occupancy."}
```

[Query] "Construct a Dynamic Scene Graph for Frame t. Analyze the force vectors applied by the instruments and the resulting tissue tension state."

```
{ "scene_event": "Blunt Dissection (Traction & Counter-traction)",
  "spatial_relations": [
    { "Grasper": "Pulls tissue [Vector: Up-Left]", "Tension": "High" },
    { "Scalpel": "Strips plane [Vector: Down-Right]", "Action": "Push" }
  ],
  "physics_reasoning": {
    "tissue_state": "Taut (Elastic limit approaching)",
    "inference": "Opposing vectors create a critical separation gap." }}
```

Figure 4. Spatial Intelligence Empowers Physics-Aware Reasoning and XR-Robotic Collaboration.

(A and B) Visualization of MedOS's spatial intelligence capabilities. (A) Depth & Spatial Parsing: The model analyzes a temporal observation window (t-1 to t+1) to resolve occlusion states, estimating the depth of the harmonic scalpel relative to the hidden Glissonian pedicle based on tissue deformation cues. (B) Dynamic Scene Graph generation: MedOS decomposes surgical maneuvers (e.g., blunt dissection) into physical force vectors (traction vs. counter-traction) and evaluates tissue tension states (e.g., Taut) to infer safety margins. (C) Examples of Counterfactual Prediction. The model simulates what-if scenarios to anticipate adverse events such as bleeding risks, margin violations, and traction errors before they materialize. (D) Quantitative recall rates for spatial intelligence tasks. MedOS significantly surpasses Gemini 3 Pro in depth parsing, spatial relation reasoning, and counterfactual prediction. (E) Application in Generative World Reconstruction (n = 1,103). Comparison of Ground Truth (left) with MedOS-generated 3D representations (center) and depth maps (right), demonstrating the platform's utility for immersive XR surgical simulation. (F) Evaluation of autonomous robotic control. The MedOS-driven robotic system demonstrates superior instrument stability compared to a junior doctor, exhibiting significantly lower metrics in static jitter, instrument drift (over 3 mins), and horizon tilt (over 10 mins). (G) Validation of real-time XR-Robot-Human Collaboration. The setup integrates MedOS livestreaming directly into XR glasses to guide robotic manipulation. The bar chart confirms that this collaborative loop enhances surgical efficiency, reducing procedure time for tasks such as Laparoscopic Cholecystectomy and Urethrovesical Anastomosis compared to unassisted human performance. The robotic and XR experiments were conducted in vitro using surgical simulators. No human patients or live animals were involved in any part of the physical experiments. The retrospective data used for model training (MedSuperVision) were sourced from de-identified, open-access educational repositories, ensuring strict privacy compliance. All photographs shown are of the authors.

Autonomous Robotic Control and XR-Enabled Human Collaboration

To translate the spatial intelligence of MedOS into physical action, we integrated the world model with a robotic surgical system to evaluate its capacity for autonomous instrument control (Figure 4F). We firstly compared the stability of a MedOS-driven robotic arm against that of a junior doctor during a laparoscopic holding task on a surgery model. The quantitative analysis focused on precision metrics including static jitter, instrument drift over three minutes, and horizon tilt over ten minutes. We observed that the AI-controlled system exhibited superior stability, maintaining significantly lower pixel deviation in jitter and drift compared to the human operator, who showed marked fluctuations due to physiological tremors. Furthermore, the robotic system maintained a balanced horizon level, whereas the junior doctor struggled with gradual tilt over time. These results further confirm that MedOS can effectively dampen the physiological inconsistencies of human motor control to achieve balanced stability in static surgical tasks.

We finally established a framework for real-time XR-Robot-Human collaboration, positioning the surgeon in a mixed reality environment where MedOS guidance is livestreamed directly to XR glasses (Figure 4G). In this setup, the human operator controls the robotic manipulators while receiving augmented visual overlays from the world model. We validated the efficacy of this collaborative system by measuring the procedure time across three distinct simulated surgeries, specifically laparoscopic cholecystectomy, urethrovesical anastomosis, and salpingostomy. The comparative data revealed that the XR-augmented cohort completed tasks significantly faster than the unassisted human group, with the most pronounced efficiency gain observed in the complex anastomosis procedure. This reduction in operative time suggests that the cognitive offloading provided by the MedOS visual guidance allows the surgeon to focus on execution rather than navigation. Ultimately, this demonstrates that synergizing human intuition with AI-driven spatial augmentation creates a surgical team that is more efficient than either entity operating alone.

DISCUSSION

The trajectory of medicine has historically bifurcated into two distinct streams: diagnostic reasoning systems that excel at data but with limited physical capability, and physical systems that possess precision but limited semantic understanding[19]. MedOS represents a shift by unifying these streams into a single embodied world model that bridges the digital and physical worlds[20,21]. Unlike standard vision-language models that treat surgical video as a passive sequence of frames [22], MedOS interprets the medical processes as a dynamic physical state space. By grounding the high-level reasoning of large language models into the egocentric reality of the surgery, we demonstrate that AI can transcend the role of a static consultant to become an active collaborator capable of guiding robotic execution [2]. This transition from AI that reads to AI that operates is critical for the next frontier of healthcare where outcomes are determined not just by the correct diagnosis but by the precision of physical intervention [23].

A central idea of MedOS is its architectural mimicry of human neurocognition through a dual-system mechanism. In surgical practice, expert performance relies on the seamless switching between expert planning and deliberate action. MedOS resolves this by decoupling trajectory planning from risk perception. The system 2 agent leverages the context of the digital world to optimize strategic workflow, while the system 1 agent operates in the physical world, executing reflex-like visual processing to flag tissue deformation or bleeding risks in real-time. This architecture ensures that the collaborative loop between surgeons, XR interfaces, and robotic agents remains synchronized within the millisecond-level constraints of surgical physics.

True surgical intelligence requires more than object recognition; it demands spatial intelligence, an understanding of depth, occlusion, and tissue mechanics. By successfully performing counterfactual predictions, such as anticipating a vessel rupture before the instrument strikes, MedOS exhibits a rudimentary form of machine intuition, potentially allowing the system to function as a safety guardrail that anticipates untoward outcomes via depth parsing. Furthermore, the success of our robotic control experiments indicates that this spatial understanding can be directly translated into motor policy, allowing the AI to dampen physiological tremors and maintain instrument stability superior to human novices.

Beyond realtime assistance, MedOS holds implications for medical education and global health equity. Medical training currently is limited by the scarcity of case volume and expert mentorship. By serving as a generative world model, MedOS can reconstruct high-fidelity digital twins from sparse video data, creating immersive environments for both human training and robotic simulation. This capability to digitize and potenrially replay the muscle memory of expert surgeons will offer a scalable solution to the shortage of medical skills. Moreover, our validation of the democratization effect suggests that a MedOS-enabled headset could provide generalist surgeons with the specialist-level guidance required to perform complex procedures safely, effectively flattening the curve of surgical proficiency.

Despite these advances, challenges remain in the translation to autonomous surgery. First, while MedOS operates with high inference speed, the latency of current XR hardware and wireless streaming can still impede the hard real-time requirements of haptic feedback loops. Second, our current world model is predominantly visual; integrating haptic sensors and force-feedback data will be necessary to achieve a complete understanding of tissue interaction. Third, the sim-to-real gap persists; while MedOS excels in predictive reasoning and collaborative control, closing the loop to allow the AI to autonomously execute complex maneuvers requires rigorous safety verification and fault-tolerant control policies. Future work will focus on integrating multimodal sensory streams and expanding the MedSuperVision benchmark to include multi-surgeon and multi-robot collaborations. Ultimately, MedOS establishes the computational foundation for the future of interventional medicine where human intuition and artificial intelligence converge to ensure the future that every patient receives expert-level care.

In summary, MedOS establishes a world model for embodied medical intelligence, bridging the gap between abstract clinical reasoning and physical surgical execution. By synergizing a dual-system cognitive architecture with a physics-aware world model, we enable AI to transcend the digital screen and actively participate in the operating room via XR-enabled human-robot collaboration. Validated on the large-scale MedSuperVision benchmark, MedOS suggests that applying expert-level diagnostic logic may facilitate better management of the dynamic complexity of surgery. Ultimately, this platform offers a scalable path toward autonomous intervention, where human-AI collaboration democratizes access to expert surgical care and push forward the boundaries of medicine.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to NVIDIA, particularly the XR/AR-VR, VLM/Agentic-AI, Edge Computing, and Business Partnership teams, for the VLM fine-tuning, agentic AI integration, and XR live streaming. We acknowledge VITURE for providing the AR/XR hardware and thank their team for technical assistance with wireless streaming and Unity/Android application support. We also express our gratitude to Dr. Ze Yuan and the UReality team for facilitating the acquisition of VR/XR hardware and for their software support. The results on cancer patient survival analysis here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Yingcheng Charles Wu thanks Dr. Qiang Gao at Fudan University for discussions. Mengdi Wang acknowledges support from by the Princeton AI Lab, Google Faculty Award.

AUTHOR CONTRIBUTIONS

L.C. and M.W. conceived this study, supervised the project, acquired funding, conceived the concept, supervised the methodology, and wrote the original draft. Y.C.W. did the methodology, agent development, data curation, visualization, and writing the original draft. M.Y. assisted with the training of the vision-language model and helped writing the draft. B.S. and Z.B. provided support for the robotic hardware integration. Z.B. supervised the robotics research. Z.Z., D.Y., X.W., Y.W., J.F., R.J., Ha.W., K.P., and K.A.Y. contributed to data collation and participated in discussions.

DECLARATION OF INTEREST

None to declare.

REFERENCES

1. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature *620*, 172-180. 10.1038/s41586-023-06291-2.
2. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature *616*, 259-265. 10.1038/s41586-023-05881-4.
3. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al. (2025). Toward expert-level medical question answering with large language models. Nat Med *31*, 943-950. 10.1038/s41591-024-03423-7.
4. Loni, M., Poursalim, F., Asadi, M., and Gharehbaghi, A. (2025). A review on generative AI models for synthetic medical text, time series, and longitudinal data. NPJ Digit Med *8*, 281. 10.1038/s41746-024-01409-w.
5. Asciak, L., Kyeremeh, J., Luo, X., Kazakidi, A., Connolly, P., Picard, F., O'Neill, K., Tsaftaris, S.A., Stewart, G.D., and Shu, W. (2025). Digital twin assisted surgery, concept, opportunities, and challenges. NPJ Digit Med *8*, 32. 10.1038/s41746-024-01413-0.
6. Marcus, H.J., Ramirez, P.T., Khan, D.Z., Layard Horsfall, H., Hanrahan, J.G., Williams, S.C., Beard, D.J., Bhat, R., Catchpole, K., Cook, A., et al. (2024). The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nat Med *30*, 61-75. 10.1038/s41591-023-02732-7.
7. Brady, O., Nulty, P., Zhang, L., Ward, T.E., and McGovern, D.P. (2025). Dual-process theory and decision-making in large language models. Nature Reviews Psychology, 1-16.
8. Bergamaschi Ganapini, M., Campbell, M., Fabiano, F., Horesh, L., Lenchner, J., Loreggia, A., Mattei, N., Rossi, F., Srivastava, B., and Venable, K. (2025). Fast, slow, and metacognitive thinking in AI. npj Artificial Intelligence *1*, 27.
9. Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., Wang, X., Feng, M., Jin, Y., and Chen, J. (2025). A survey of embodied ai in healthcare: Techniques, applications, and opportunities. arXiv preprint arXiv:2501.07468.
10. Yao, Z., Zhang, Z., Tang, C., Bian, X., Zhao, Y., Yang, Z., Wang, J., Zhou, H., Jang, W.S., and Ouyang, F. (2024). Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. arXiv preprint arXiv:2410.01553.
11. Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J., Michael, J., and Bowman, S.R. (2024). Gpqa: A graduate-level google-proof q&a benchmark.
12. Zhao, B., Zhang, X., Chen, M., and Wang, Y. (2023). A real-world data analysis of acetylsalicylic acid in FDA Adverse Event Reporting System (FAERS) database. Expert opinion on drug metabolism & toxicology *19*, 381-387.
13. Hutter, C., and Zenklusen, J.C. (2018). The cancer genome atlas: creating lasting value beyond its data. Cell *173*, 283-285.
14. Shi, C., Liu, S., Tian, X., Wang, X., and Gao, P. (2021). A TP53 mutation model for the prediction of prognosis and therapeutic responses in head and neck squamous cell carcinoma. BMC Cancer *21*, 1035. 10.1186/s12885-021-08765-w.
15. Wang, C., Sandhu, J., Tsao, A., and Fakih, M. (2022). Presence of Concurrent TP53 Mutations Is Necessary to Predict Poor Outcomes within the SMAD4 Mutated Subgroup of Metastatic Colorectal Cancer. Cancers (Basel) *14*. 10.3390/cancers14153644.
16. Chen, Z., Luo, Z., Zhang, D., Li, H., Liu, X., Zhu, K., Zhang, H., Wang, Z., Zhou, P., and Ren, J. (2023). TIGER: a web portal of tumor immunotherapy gene expression resource. Genomics, proteomics & bioinformatics *21*, 337-348.
17. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., and Wu, Y. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.
18. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., and Lv, C. (2025). Qwen3 technical report. arXiv preprint arXiv:2505.09388.
19. Schmidgall, S., Opfermann, J.D., Kim, J.W., and Krieger, A. (2025). Will your next surgeon be a robot? Autonomy and AI in robotic surgery. Sci Robot *10*, eadt0187. 10.1126/scirobotics.adt0187.
20. Long, Y., Lin, A., Kwok, D.H.C., Zhang, L., Yang, Z., Shi, K., Song, L., Fu, J., Lin, H., Wei, W., et al. (2025). Surgical

embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. Sci Robot *10*, eadt3093. 10.1126/scirobotics.adt3093.

21. Shademan, A., Decker, R.S., Opfermann, J.D., Leonard, S., Krieger, A., and Kim, P.C. (2016). Supervised autonomous robotic soft tissue surgery. Sci Transl Med *8*, 337ra364. 10.1126/scitranslmed.aad9398.

22. Kiyasseh, D., Ma, R., Haque, T.F., Miles, B.J., Wagner, C., Donoho, D.A., Anandkumar, A., and Hung, A.J. (2023). A vision transformer for decoding surgeon activity from surgical videos. Nat Biomed Eng *7*, 780-796. 10.1038/s41551-023-01010-8.

23. Saeidi, H., Opfermann, J.D., Kam, M., Wei, S., Leonard, S., Hsieh, M.H., Kang, J.U., and Krieger, A. (2022). Autonomous robotic laparoscopic surgery for intestinal anastomosis. Sci Robot *7*, eabj2908. 10.1126/scirobotics.abj2908.

METHODS

MedOS System Architecture and Core Agents

The MedOS framework functions as a unified agentic world model designed to bridge the historical divide between abstract clinical reasoning and physical surgical execution. As illustrated in Figure 1A, the architecture operates across two fundamental planes, the Digital World and the Physical World, to align medical logic with surgical reality. The system integrates a multi-agent workflow orchestrated by a Coordinator Agent that decomposes complex queries for specialized modules, ensuring that high-level strategic planning translates effectively into real-time intervention. By grounding abstract knowledge into a dynamic state space, the platform enables the AI to function not merely as a consultant but as a perceiving co-physician capable of active collaboration. Below, the roles of specific agents and the dual-system cognitive architecture are detailed to illustrate their collective function.

Dual-System Cognitive Architecture To master the high-stakes dynamics of interventional medicine, MedOS mimics expert human cognition through a dual-system mechanism. This architecture operationalizes the seamless toggling between deliberate strategy and reflexive action observed in clinical practice. The System 1 Fast module processes real-time egocentric video streams to handle millisecond-level risk perception and reflex-like guidance. Upon perceiving adverse states such as fibrotic adhesions, it immediately reasons about tissue friability and guides robotic action to specific interventions like suction dissection. Conversely, the System 2 Slow module coordinates high-level planning and trajectory optimization based on the full digital context. It processes macro-context such as patient demographics and meso-context like perioperative plans to ensure that immediate actions align with long-term clinical goals.

Coordinator and Specialized Agents The Coordinator Agent serves as the central orchestration node, managing the workflow by breaking down complex clinical queries into structured sub-tasks. It distributes these tasks to a suite of specialized agents, including an EHR Agent for longitudinal history, a Guideline Agent for standard of care, a Radiology Agent for imaging analysis, and a Pathology Agent for histological data. This distributed processing ensures that all dimensions of a clinical case are evaluated with domain-specific rigor.

Reasoning and Critic Agents The core Reasoning Agent executes a structured thinking template driven by evidence synthesis and causal inference. It integrates outputs from specialized agents to formulate cohesive clinical strategies. To ensure safety and reliability, a Self-evolving Critic Agent continuously evaluates these plans. This agent functions as a governance mechanism, capable of rejecting unsafe proposals, issuing warnings, or approving valid strategies, thereby creating a robust quality control loop within the reasoning process.

The MedOS Agentic Tool System MedOS integrates a diverse array of digital and physical tools to facilitate the transition from information retrieval to physical action. The system employs a tool ocean that provides capabilities ranging from information synthesis to spatial reconstruction.

Digital Logic Tools To support the reasoning process in the Digital World (Level 1), the system utilizes specialized computational tools. These include PubMed_Search for literature retrieval, FDA_Drug for pharmacovigilance data, and a Surgical_Risk_Calculator for quantitative risk assessment. For longitudinal analysis, the system employs tools to process lifelong tokens, identifying patient phenotypes such as cirrhosis and risks like portal hypertension. This suite allows the system to establish a strategic baseline grounded in comprehensive clinical evidence.

Physical Perception and Control Interface In the Physical World (Level 2), MedOS models a 3D state space that includes the egocentric view, real-time scene depth, and instrument interactions. The Anatomy_Recon_Tool is deployed to decode visual input from extended reality glasses, enabling the system to execute counterfactual prediction. To translate this spatial intelligence into action, the system interfaces with high-bandwidth XR Streaming and Robotic Control modules. This allows MedOS to simulate a physics model where it can reason and predict tissue features, effectively linking digital analysis to physical instrument control.

Dual-System Learning and Self-Evolution

We demonstrate that the MedOS world model possesses capabilities for both training-time optimization and inference-time scaling. This allows the system to evolve its performance for both reflexive physical tasks and complex cognitive reasoning.

Dual-System Training Strategy To distinctively optimize the two sub-modules of our cognitive architecture, we employed a training strategy combining Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO) based on Qwen3-VL-8B-Instruct. For the System 1 Fast module, the GRPO objective focused on immediate next action prediction and risk detection, optimizing the model for high-frequency visual processing. For the System 2 Slow module, the optimization targeted trajectory planning and chain-of-thought reasoning. This bifurcated training approach ensures that the model achieves state-of-the-art performance in both dynamic physical tasks and abstract logical reasoning.

Model Specification and Training Recipe We utilized Qwen3-VL-8B-Instruct as the backbone. The dual-system architecture is implemented via two distinct Low-Rank Adaptation (LoRA) modules tailored for different distinct objectives: the System 1 module is optimized for latency-sensitive object detection and action classification, while the System 2 module employs Group Relative Policy Optimization (GRPO) to reward long-chain reasoning steps. The physics model is computationally defined as a video prediction head that generates future frames (t+1) conditioned on instrument force vectors. This allows the model to simulate tissue deformation (e.g., stretching, tearing) in latent space before executing actions.

Inference-Time Scaling Beyond static training, MedOS demonstrates inference-time scaling properties. By increasing the token budget allocated for the System 2 thinking process from 1x to 9x, we observed a systematic improvement in model performance. This positive correlation provides direct evidence that the dual-system design enables the AI to think harder and evolve its strategies for complex clinical scenarios, effectively adapting its computational effort to the difficulty of the task at hand.

Benchmark Design, Baselines, and Evaluation Methods

Construction of MedSuperVision Dataset To enable the AI to perceive the physical reality of surgery beyond static frames, we manually assembled MedSuperVision (MSV), a large-scale, expert-annotated surgical video dataset. We devised a rigorous four-phase curation protocol. Phase I involved the aggregation of diverse egocentric videos enriched by expert narratives to capture intent. Phase II processed this footage through time-frame segmentation (t-1, t, t+1) by 10 seconds, chain-of-thought extraction, and patient health information removal. The resulting dataset contains 85,398 minutes of video spanning multiple disciplines including hepatobiliary, gastrointestinal, urology, vascular, and thoracic procedures. All videos feature narrations by 1,882 clinical experts in the educational surgery videos. The dataset was split into 80% for training and 20% for validation.

Reasoning Evaluation To validate the reasoning capabilities of the system, we utilized challenging biomedical benchmarks. On MedQA (USMLE), MedOS was evaluated against frontier models including Gemini 3 Pro, GPT-5.2 Thinking, and Claude 4.5 Opus. Similarly, performance was assessed on the GPQA benchmark for expert-level reasoning. We further extended the evaluation to autonomous clinical research tasks, quantifying the system's ability to execute workflows such as meta-analysis of FAERS data and survival analysis of TCGA genomic cohorts.

Spatial Intelligence Evaluation To evaluate physical perception, we designed specific tasks to test spatial intelligence. Depth & Spatial Parsing involved analyzing a temporal observation window to resolve occlusion states and estimate the relative depth of instruments. Spatial Relation Reasoning required the model to decompose complex maneuvers like blunt dissection into force vectors and tissue tension states. Counterfactual Prediction tested the model's ability to foresee adverse events such as bleeding risks or margin violations before they materialized. Performance was measured using recall rates and compared against the baseline Gemini 3 Pro model.

Human Expert Rating To validate clinical utility, we conducted a blinded human expert rating composed of 5 licensed medical doctors. These experts rated 100 surgical scenes processed by both MedOS and Gemini 3 Pro. The win rate was calculated to determine the preference for MedOS-generated guidance in real-world surgical interpretation.

Experimental Validation of Physical Interaction

Robotic Control Validation To translate spatial intelligence into physical action, we integrated the world model with a robotic surgical system. We compared the stability of a MedOS-driven robotic arm against that of a junior doctor during a laparoscopic holding task. The quantitative analysis focused on precision metrics including static jitter (pixel deviation), instrument drift over three minutes, and horizon tilt over ten minutes. This comparison aimed to verify the system's ability to dampen physiological inconsistencies and maintain instrument stability.

XR-Enabled Human Collaboration Study We established a framework for real-time XR-Robot-Human collaboration, positioning the surgeon in a mixed reality environment where MedOS guidance is livestreamed directly to XR glasses. To evaluate the efficacy of this system, we measured the procedure time across three distinct simulated surgeries: laparoscopic cholecystectomy, urethrovesical anastomosis, and salpingostomy. We compared the performance of an XR-augmented cohort against an unassisted human group to quantify efficiency gains provided by the collaborative loop.

Human-AI Collaboration Study To evaluate the system's capacity to democratize expertise, we designed a study involving participants (n = 24) with varying levels of hierarchy, educational backgrounds, and physiological states. Participants included registered nurses, medical students, resident physicians, and attending physicians. We measured diagnostic accuracy on MedQA tasks under unaided and AI-augmented conditions. We specifically investigated the system's impact on mitigating cognitive deficits in sleep-deprived post-call physicians and closing educational gaps between graduates of top-tier and unranked medical schools.

Open-access patient data collection This study utilizes de-identified data from open-access educational repositories. PHI Removal: We implemented a privacy pipeline. For the videos that patient appears, we blurred their faces and text. Chain-of-Thought Extraction: Expert narratives were transcribed and extracted using Qwen, and Gemini-3-Pro was employed to structure these unstructured narrations into formal step-by-step reasoning traces, which were subsequently verified by the licenced doctor for accuracy.

Human Participation The human-AI collaboration component was conducted in a simulated environment to evaluate software performance. This study involves participants composed of doctors, nurses, and medical students for the purpose of evaluating AI model performance (via exam question answering) and subjective quality rating. Written/Oral informed consent was obtained from all participants. The study was conducted in adherence to the Declaration of Helsinki. The robotic and XR experiments were conducted in vitro using surgical simulators. No human patients or animals were involved in the experiments. The MedSuperVision dataset was constructed exclusively from publicly available, open-access educational resources. All data were retrospective and fully de-identified prior to analysis. No direct interaction with patients occurred, and no private health information (PHI) was accessed or utilized. All photographs shown are of the authors. In accordance with regulations 45 CFR 46, this portion of the study does not qualify as human subjects research requiring institutional review board (IRB) oversight.

Statistical Analysis

For genomic oncology analysis, we utilized Kaplan-Meier curves to visualize survival probabilities and the Log-rank test to determine statistical significance ($P < 0.001$) between co-mutation and wild-type cohorts. In the meta-analysis of clinical trial data, forest plots were generated to visualize differences in inflammatory factor levels. All comparisons in the human-AI collaboration study were quantified by calculating the percentage improvement in accuracy scores across the defined cohorts.

Data and Code Availability

The MedSuperVision dataset, agentic framework, and training scripts will be available upon request (https://forms.gle/oWJwuri18y4rRuAb8).