Surgery AI: Multimodal Process Mining and Mixed Reality for Real-time Surgical Conformance Checking and Guidance

Aleksandar Gavric, Dominik Bork and Henderik A. Proper

Business Informatics, TU Wien, Erzherzog-Johann-Platz 1, Vienna, 1040, Austria

Abstract

This paper discusses an end-to-end methodology for real-time surgical conformance checking that uses multimodal process mining, mixed reality (MR), and large language model (LLM) prompting. Our approach aims to support surgeons and medical teams by comparing *as-is* operational data—captured through a variety of sensors including MR-based gaze tracking—with a reference surgical process model encoded in Business Process Modeling Notation (BPMN). We illustrate how *shallow* and *deep* human-in-the-loop feedback mechanisms can be integrated with chain-of-thought prompting to provide relevant, context-aware, and iterative feedback during surgery. We further indicate which aspects of the surgery can be monitored (and hence queried) by our multimodal process mining engine. By enabling precise, actionable feedback during critical surgical procedures, our approach enhances the ability to identify deviations, ensure adherence to best practices, and reduce human error. Ultimately, this methodology empowers surgical teams to make data-driven adjustments, promotes better patient outcomes, and allows hospitals to monitor surgical conformance effectively, setting a new standard for process-driven healthcare assistance.

Keywords

Multimodal data analysis, Mixed reality, Surgery AI, Surgical guidance, Process mining, BPMN, LLM, Healthcare, assistant

1. Introduction

Modern surgical procedures are intricate and involve numerous steps, actors, instruments, and real-time decisions. Ensuring that each step in the *as-is* surgery conforms to a reference (or "desired") model is crucial for patient safety, consistent outcomes, and compliance with institutional guidelines. Traditional methods of process oversight often rely on paper-based checklists or single-modality digital signals (e.g., time stamps of major milestones), which offer limited real-time insight.

Process mining [1] addresses this gap by extracting event logs from complex systems and reconstructing an *as-is* process model. Yet standard process mining may overlook the depth of real-time information available from modern medical devices, images, sensor data, and user interactions in an operating room [2]. The growing accessibility of mixed reality (MR) systems and advanced wearable sensors (like gaze trackers) opens the door to *multimodal*

✉ aleksandar.gavric@tuwien.ac.at (A. Gavric); dominik.bork@tuwien.ac.at (D. Bork);
henderik.proper@tuwien.ac.at (H. A. Proper)
0009-0005-1243-7722 (A. Gavric); 0000-0001-8259-2297 (D. Bork); 0000-0002-7318-2496 (H. A. Proper)

process mining [2, 3], where we capture a richer set of signals beyond textual or numeric logs (e.g., surgeon gaze, instrument position, physical environment changes, real-time vitals).

Meanwhile, Large Language Models (LLMs) allow us to harness *conversational* and *chain-of-thought* prompting to incorporate human expertise dynamically. Surgeons, nurses, and other staff can interact with the system at various depths: (1) *Shallow feedback*: Quick confirmations or corrections to immediate queries (e.g., "Is the incision completed?"), and (2) *Deep feedback*: More reflective input that leads to refining the underlying process model or augmenting the system's domain knowledge [4].

Mixed reality interfaces can further project relevant information in the surgical environment, supporting *Spatial Conceptual Modeling* [5] to visualize conformance data in situ. This integration bridges the gap between human expertise and automated systems by enabling real-time contextual feedback and adaptive process modeling. For instance, visual overlays or auditory alerts can notify surgeons of deviations from standard procedures or highlight critical decision points, leveraging AI-based interpretation of multimodal data [6].

2. Related Work

The integration of artificial intelligence (AI) and mixed reality (MR) in surgical environments has emerged as a promising research area, driven by advancements in computer vision, language models, and multimodal process mining. This section reviews the most relevant contributions in this domain.

Recent efforts, such as Surgical-LLaVA [7], have demonstrated the potential of large language and vision models for understanding surgical scenarios, offering a foundation for enhanced decision support systems. Similarly, Yuan et al. [8] proposed a procedure-aware surgical video-language pretraining approach, utilizing hierarchical knowledge augmentation to improve the interpretability of surgical workflows. Digital twins, as described by Ding et al. [9], provide a unifying framework for surgical data science, leveraging geometric scene understanding to create comprehensive models of the operating room (OR). Complementing this, holistic OR domain modeling using semantic scene graphs has been explored by Özsoy et al. [10], enabling a detailed representation of surgical environments.

Further advancements in surgical scene graph knowledge have been achieved by Yuan et al. [11], who incorporated scene graphs into visual question answering (VQA) systems for surgical applications, thereby enhancing context-awareness in automated systems. The Ophnet benchmark by Hu et al. [12] provides a large-scale video dataset for ophthalmic surgical workflow understanding, facilitating the development of robust AI models in the domain.

Incorporating mixed reality into surgical planning and execution has also gained traction. Bracale et al. [13] highlighted the utility of MR in preoperative planning for colorectal surgery, showcasing its potential to improve surgical outcomes. From a conceptual perspective, Fill [5] introduced spatial conceptual modeling, which anchors knowledge in the physical world using augmented reality technologies, enabling innovative applications in medical and other domains.

Our prior contributions have laid the groundwork for advancing multimodal process mining and its applications. In Multimodal Process Mining [2], we introduced an approach to enrich traditional process mining with multimodal evidence, capturing data from diverse sources such as sensors, images, and user interactions. Building on this, we explored how to enhance business process event logs with multimodal evidence in [3], demonstrating the potential for deeper insights. In [4], we addressed the challenge of tailoring multimodal data representations to stakeholder-specific terminology for improved interpretability. Finally, in [6], we extended the multimodal paradigm to conceptual modeling, showcasing how AI can leverage visual and auditory cues to interpret UML diagrams. These contributions collectively highlight the potential of multimodal approaches in augmenting traditional process and conceptual modeling practices.

By uniting algorithmic-symbolic rigor with LLM-driven sub-symbolic flexibility and human expertise, our approach transcends the constraints of rule-based process mining, enabling a more dynamic and contextually rich analysis of surgical workflows.

3. Methodology Overview

We formalize the multimodal process monitoring and adaptive feedback mechanism as an optimization problem over a hybrid state space $\mathcal{S}$ consisting of structured process models, multimodal sensor inputs, and user feedback mechanisms.

State Representation Let the state at time t be represented as: $S_t = (M_t, X_t, U_t)$, where $M_t \in \mathcal{M}$ represents the current process model state (e.g., BPMN graph representation, stored in a Retrieval Augmented Graph [14]), $X_t \in \mathcal{X}$ denotes the vector of multimodal sensor observations (e.g., gaze tracking, instrument logs, voice commands), and $U_t \in \mathcal{U}$ captures the human feedback at time t, either shallow (e.g., confirmation) or deep (e.g., structural model changes).

Transition Function The state transition function $T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ maps the current state and action to a new state, $S_{t+1} = T(S_t, A_t)$ where A_t represents an action taken by the system or user, such as:

- A_t^S (System Actions): Process conformance checking, real-time alerting, adaptive workflow modification,
- A_t^H (Human Actions): Explicit feedback confirmation, model refinement, procedural adjustments.

Objective Function The system aims to minimize a cumulative deviation function J that quantifies non-conformance with the desired process model while maximizing the incorporation of human feedback. This ensures continuous process adaptation and human-in-the-loop refinement over time.

3.1. Application to a Specific Use Case: Surgery

We instantiate our proposed framework in the context of surgery, a domain characterized by strict procedural adherence and real-time decision-making.

Process Modeling and Sensor Integration The BPMN model for procedures includes predefined steps such as incision, trocar placement, laparoscope insertion, and organ manipulation. The system continuously maps real-world observations to this structured model through:

- Vision-based Instrument Detection (X_t^{inst}): Identifies tool usage and compares with expected sequences.
- Eye-tracking (X_t^{gaze}): Confirms if surgeons are focusing on critical areas at appropriate steps.
- Hand Gesture Recognition (X_t^{gest}): Detects compliance with required movements (e.g., correct suturing technique).
- Voice Commands (X_t^{voice}): Captures surgeon-nurse communications for validation.
- Real-time Imaging (X_t^{img}): Analyzes anatomical landmarks for correct procedure execution.

Illustrative Scenario Consider a scenario where a surgeon employs a novel technique requiring a secondary incision. The system detects a deviation (X_t^{img} and X_t^{inst} differ from the expected process).

Figure 1 provides a high-level schematic overview of our proposed framework adapted for the domain of surgery. Two major phases of human-in-the-loop involvement are depicted:

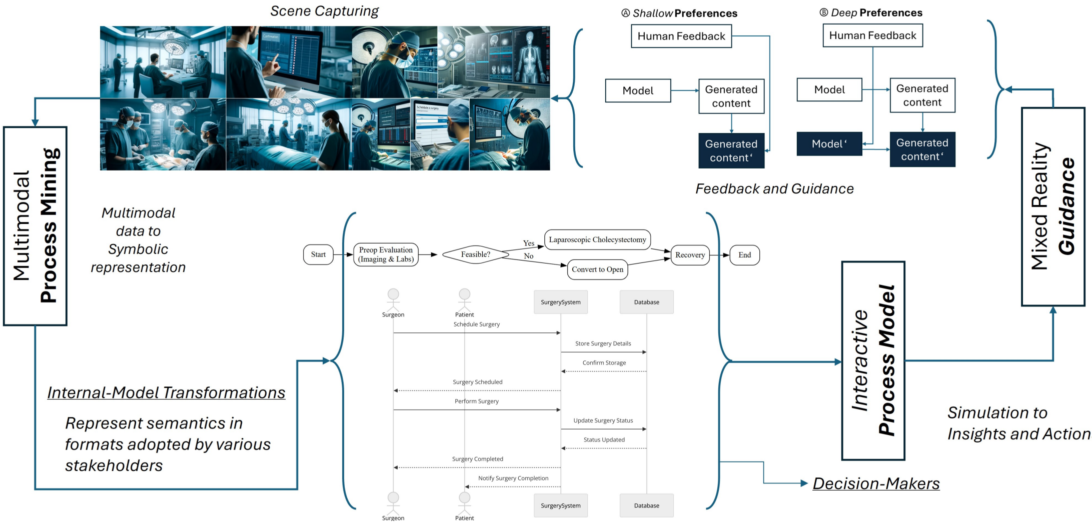

Figure 1: (Top) Illustration of human feedback via (A) Shallow and (B) Deep approaches, and a multimodal scene tracking. (Bottom) The solution's pipeline and illustrated (simplified) conceptual (process) model representation adjusted for conformance checking and guidance.

1. **Shallow Feedback (A):**
    - The system continuously captures data from multiple sources (e.g., gaze tracking, instrument usage, sensor logs).
    - It compares the *as-is* process to the *desired* BPMN model.

- When a potential discrepancy or question arises, the system prompts the user (surgeon, nurse, etc.) for feedback.
- The user provides confirmation, correction, or small clarifications. This feedback is used to adjust or annotate the *current* run-time process instance.

2. **Deep Feedback (B):**
    - In-depth reflections by human experts are used to refine the *model* itself or the methods that interpret the captured data.
    - For instance, if the current process model does not account for a new device or step introduced by the surgical team, deep feedback cycles can lead to an updated BPMN model or a reconfiguration of the data capture pipeline.
    - Over time, repeated deep feedback loops result in an evolving knowledge base that is more robust and better tailored to each specific surgery or environment.

A critical component of our setup is the Mixed Reality (MR) environment, which serves multiple purposes:

- **Precision of Multimodal Recordings:** By using MR headsets, the system can track the surgeon's gaze in relation to specific instruments or areas of the patient's body. Likewise, position and orientation of surgical staff can be recorded.
- **Spatial Conceptual Modeling for Feedback:** We build on *Spatial Conceptual Modeling* [5], which allows us to overlay real-time process conformance data directly into the OR environment. For example, a soft highlight (visible in the MR headset) might appear over the next instrument to be used, or an alert icon might appear above a piece of equipment that must be sanitized.

3.2. Chain-of-Thought LLM Prompting

The proposed methodology employs a *conversation engine* powered by Large Language Models (e.g., GPT variants) that can: (1) **Parse sensor events** and interpret them in the context of the BPMN model, (2) **Generate feedback prompts** when conformance might be violated, (3) **Solicit clarifications and deeper insights** from the surgeon or nurse (for refining the model), and (4) **Provide intermediate "food-for-thought"** (chain-of-thought) to guide the surgical team or system designers on why certain steps are suggested or flagged.

A key component of our methodology is the construction of *well-curated LLM prompts* that merge domain knowledge (e.g., typical steps in a laparoscopic procedure) with real-time sensor data (e.g., the last tool recognized by a vision sensor was a cauterizing instrument). Below is a conceptual example of the layered prompts:

Example Prompt for Understanding Mixed Reality Inputs

> **System (LLM context):** *"The surgeon's gaze has been fixated on the laparoscope for 5 seconds, and the nurse passed the laparoscope 10 seconds ago. The BPMN model indicates we are in the "Insert Laparoscope" task. Confirm if this step is complete. If uncertain, ask for feedback from the user."*

Example Prompt for Generating Feedback

> **System (LLM context after receiving user input):** *"User indicated that they are testing a new technique requiring a secondary incision. The current BPMN model does not include this step. Rather than adding an optional sub-process, this should be modeled as an alternative process path. Insert an exclusive Gateway with the existing technique subprocess and the new technique subprocess as subsequent elements. Record new recommended tasks accordingly. Provide a revised BPMN snippet."*

3.3. Aspects to Monitor for Multimodal Process Mining

Beyond the questions a surgeon might explicitly ask, the system continuously mines data to update the *as-is* process model. Table 1 outlines various aspects of surgery that are relevant for conformance checking, each corresponding to multimodal sensor inputs.

Table 1: Aspects to Monitor for Real-Time Surgical Conformance Checking

Aspect	Sensor/Data Source	Reason for Relevance to Conformance
Instrument Usage	Instrument detection via computer vision (camera) + staff input logs	To confirm correct usage sequence, detect missing/extra usage, alert if an instrument wasn't sterilized, etc.
Gaze Tracking	MR headset with eye-tracking	To assess if the surgeon is focused on the correct region/patient area. Non-conformance might arise if the surgeon fails to visually confirm a step (e.g., lack of inspection).
Hand Gestures	MR/IR sensors, glove-based trackers	To detect if certain steps (e.g., suturing technique) are performed in standard manner, or to confirm that a gesture-based command has been recognized.
Patient Vitals	Anesthesia machine logs, heart rate monitor, SpO2 sensor	To ensure anesthesia compliance steps, watch for anomalies that might require altering the process (e.g., emergency protocols).
Tool Count	Vision-based object detection, manual logs from nurses	To check if the correct number of instruments/sponges are present before closure (avoid retained surgical items).
Environment Sterility	UV sensor logs, staff compliance logs (handwashing, glove changes)	Conformance checking for infection control steps, verifying that each area is sanitized prior to the next step.

(continued on next page)

...(table continued)

Aspect	Sensor/Data Source	Reason for Relevance to Conformance
Surgeon/Nurse Position	MR device tracking (position/orientation)	To ensure correct posture or vantage point is taken for certain steps (e.g., for laparoscopic approach, a specific angle might be recommended).
Incision and Wound	Camera feed from laparoscope or overhead camera	To verify compliance with recommended incision size, location, and closure technique.
Anatomical Landmarks	Imaging data (e.g., real-time ultrasound, MRI overlays)	To confirm that the correct organ or region is identified before proceeding (e.g., right kidney instead of left).
Timeline / Timing	Digital clock + event logs	To confirm that each task is within an acceptable time window (e.g., prophylactic antibiotics repeated in time).
Communication Logs	Voice recognition or typed notes	To verify that critical verbal confirmations are done (e.g., "Time Out" procedure).
Clinical Documentation	EHR (Electronic Health Record) system	To confirm data entry is complete and consistent with the surgical plan (e.g., procedure codes, lab results).
Unexpected Events	Automatic anomaly detection (vitals, sudden camera movements)	To trigger re-routing of the BPMN process to an emergency sub-process if necessary (e.g., severe hemorrhage).
Surgeon/Staff Vitals	Smartwatches, wearable health trackers	To monitor the physical state of surgeons and staff (e.g., heart rate, stress levels, fatigue) and proactively suggest breaks or duty switches when signs of exhaustion or stress are detected, especially in surgeries involving multiple surgeons.

4. Proposed Evaluation and Future Work

We propose a multi-faceted evaluation framework, leveraging established surgical video datasets [15, 16, 17] to benchmark performance across several key metrics:

- **Annotation Accuracy:** Measure tool and event recognition accuracy against expert annotations.
- **Temporal Consistency:** Evaluate the alignment between detected events and ground truth timelines, ensuring timely alerts and correct sequencing.
- **Process Conformance:** Assess the system's ability to detect deviations from standard

protocols using conformance checking metrics, such as deviation frequency and critical event misclassifications.

- For evaluating the **robustness across modalities**, we will analyze performance consistency across different sensor inputs to ensure reliable multimodal integration.

By applying these metrics on diverse datasets from cataract [15], laparoscopic [16], and robotic surgery domains [17], we aim to demonstrate the system's versatility and readiness for real-time surgical support.

Our roadmap for future work outlines key steps to ensure continuous improvement and user-centric development: (1) Extend evaluations to large, diverse datasets, including complex and rare surgical procedures, (2) implement rigorous testing on annotated datasets to validate real-time performance and scalability, and (3) engage with final users (surgeons, nurses) to gather feedback on system performance and usability.

5. Conclusion

By integrating *multimodal process mining* with *Mixed Reality* and *LLM-driven chain-of-thought prompting*, we propose a highly granular, real-time conformance checking methodology for surgical processes. User confirmations can augment immediate decisions in the operating room, while deeper reflection iteratively improves the process model over time.

As a result, surgeons can rely on the system to (1) provide step-by-step prompts and clarifications, (2) alert them when tasks are out of sequence or incomplete, (3) suggest new tasks when a procedure deviates from established protocols, and (4) support advanced analytics using chain-of-thought reasoning that ties sensor data to context-specific knowledge of surgical procedures.

Despite its promising capabilities, our approach has several limitations. Inaccuracies in sensor data (e.g., video feeds, gaze tracking) or inconsistent data quality may affect the system's reliability. The methodology validated on selected surgical datasets, may require significant adaptation to perform effectively across diverse surgical procedures and environments. Achieving true real-time performance can be challenging due to the computational complexity of multimodal data fusion and chain-of-thought processing.

Addressing these threats and limitations through continued testing, iterative user feedback, and technological refinements will be essential for future deployments in dynamic surgical environments.

References

[1] W. M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer, Berlin, Heidelberg, 2011. URL: https://doi.org/10.1007/978-3-642-19345-3. doi:10.1007/978-3-642-19345-3.

[2] A. Gavric, D. Bork, H. Proper, Multimodal process mining, in: 26th International Conference on Business Informatics (CBI), 2024.

[3] A. Gavric, D. Bork, H. Proper, Enriching business process event logs with multimodal evidence, in: The 17th IFIP WG 8.1 Working Conference on the Practice of Enterpris Modeling (PoEM), 2024.

[4] A. Gavric, D. Bork, H. Proper, Stakeholder-specific jargon-based representation of multimodal data within business process, in: Companion Proceedings of the 17th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling (PoEM Forum 2024), 2024.

[5] H.-G. Fill, Spatial Conceptual Modeling: Anchoring Knowledge in the Real World, Springer Nature Switzerland, Cham, 2024, pp. 35–50. URL: https://doi.org/10.1007/978-3-031-56862-6_3. doi:10.1007/978-3-031-56862-6_3.

[6] A. Gavric, D. Bork, H. Proper, How does uml look and sound? using ai to interpret uml diagrams through multimodal evidence, in: 43rd International Conference on Conceptual Modeling (ER), 2024.

[7] J. Jin, C. W. Jeong, Surgical-llava: Toward surgical scenario understanding via large language and vision models, 2024. URL: https://arxiv.org/abs/2410.09750. arXiv:2410.09750.

[8] K. Yuan, V. Srivastav, N. Navab, N. Padoy, Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation, arXiv preprint arXiv:2410.00263 (2024).

[9] H. Ding, L. Seenivasan, B. D. Killeen, S. M. Cho, M. Unberath, Digital twins as a unifying framework for surgical data science: the enabling role of geometric scene understanding, Artificial Intelligence Surgery 4 (2024) 109–138.

[10] E. Özsoy, T. Czempiel, E. P. Örnek, U. Eck, F. Tombari, N. Navab, Holistic or domain modeling: a semantic scene graph approach, International Journal of Computer Assisted Radiology and Surgery 19 (2024) 791–799.

[11] K. Yuan, M. Kattel, J. L. Lavanchy, N. Navab, V. Srivastav, N. Padoy, Advancing surgical vqa with scene graph knowledge, International Journal of Computer Assisted Radiology and Surgery (2024) 1–9.

[12] M. Hu, P. Xia, L. Wang, S. Yan, F. Tang, Z. Xu, Y. Luo, K. Song, J. Leitner, X. Cheng, et al., Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding, in: European Conference on Computer Vision, Springer, 2025, pp. 481–500.

[13] U. Bracale, B. Iacone, A. Tedesco, A. Gargiulo, M. M. Di Nuzzo, D. Sannino, S. Tramontano, F. Corcione, The use of mixed reality in the preoperative planning of colorectal surgery: Preliminary experience with a narrative review, Cirugía Española (English Edition) (2024).

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

[15] H. Al Hajj, M. Lamard, P.-H. Conze, S. Roychowdhury, X. Hu, G. Maršalkaitė, O. Zisimopoulos, M. A. Dedmari, F. Zhao, J. Prellberg, M. Sahu, A. Galdran, T. Araújo, D. M. Vo, C. Panda, N. Dahiya, S. Kondo, Z. Bian, A. Vahdat, J. Bialopetravičius, E. Flouty, C. Qiu, S. Dill, A. Mukhopadhyay, P. Costa, G. Aresta, S. Ramamurthy, S.-W. Lee, A. Campilho, S. Zachow, S. Xia, S. Conjeti, D. Stoyanov, J. Armaitis, P.-A. Heng, W. G. Macready, B. Cochener, G. Quellec, Cataracts: Challenge on automatic tool annotation for cataract surgery, Medical Image Analysis 52 (2019) 24–41. doi:https://doi.org/10.1016/j.media.2018.11.008.

[16] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, N. Padoy, Endonet: A deep architecture for recognition tasks on laparoscopic videos, 2016. arXiv:1602.03012.

[17] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, A. Kori, V. Alex, G. Krishnamurthi, D. Rauber, R. Mendel, C. Palm, S. Bano, G. Saibro, C.-S. Shih, H.-A. Chiang, J. Zhuang, J. Yang, V. Iglovikov, A. Dobrenkii, M. Reddiboina, A. Reddy, X. Liu, C. Gao, M. Unberath, M. Kim, C. Kim, C. Kim, H. Kim, G. Lee, I. Ullah, M. Luna, S. H. Park, M. Azizian, D. Stoyanov, L. Maier-Hein, S. Speidel, 2018 robotic scene segmentation challenge, 2020. arXiv:2001.11190.