MIDAS: A MULTIMODAL DATA ACQUISITION SYSTEM AND DATASET FOR ROBOT-ASSISTED MINIMALLY INVASIVE SURGERY

A PREPRINT

Keshara Weerasinghe[1], **Seyed Hamid Reza Roodabeh**[1], **Andrew Hawkins (MD)**[2],
Zhaomeng Zhang[1], **Zachary Schrader**[1], **and Homa Alemzadeh**[1]
[1]School of Engineering and Applied Science, University of Virginia
[2]Thoracic and Cardiovascular Surgery, Department of Surgery, University of Virginia Health System
[1]`{cjh9fw, ydq9ag, sus7sv, zgr3jk, ha4d}@virginia.edu`
[2]`adh2tp@uvahealth.org`

March 9, 2026

ABSTRACT

Background: Robot-assisted minimally invasive surgery (RMIS) research increasingly relies on multimodal data, yet access to proprietary robot telemetry remains a major barrier. We introduce MiDAS, an open-source, platform-agnostic system enabling time-synchronized, non-invasive multimodal data acquisition across surgical robotic platforms.

Methods: MiDAS integrates electromagnetic and RGB-D hand tracking, foot pedal sensing, and surgical video capturing without requiring proprietary robot interfaces. We validated MiDAS on the open-source Raven-II and the clinical da Vinci Xi by collecting multimodal datasets of peg transfer and hernia repair suturing tasks performed by surgical residents. Correlation analysis and downstream gesture recognition experiments were conducted.

Results: External hand and foot sensing closely approximated internal robot kinematics, and non-invasive motion signals achieved gesture recognition performance comparable to robot kinematics.

Conclusions: MiDAS enables reproducible multimodal RMIS data collection and is released with annotated datasets, including the first multimodal dataset capturing hernia repair suturing on high-fidelity simulation models.

Code & Dataset: `https://uva-dsa.github.io/MiDAS/`

Keywords Robot-Assisted Minimally Invasive Surgery · Surgical Robotics · Multimodal Surgical Data Acquisition · Non-invasive Motion Sensing · Surgical Gesture Recognition · Surgical Activity Datasets

1 Introduction

Robot-assisted surgery (RAS) has become central to minimally invasive surgery, yet surgical outcomes remain heavily dependent on surgeon skill Wang and Fey [2018]. The multimodal data generated during RAS, including stereo video, robot kinematics, and interaction events offer unprecedented opportunities for training assessment, error detection and skill transfer research Hung et al. [2018].

At the system level, a surgeon operates from a master console with a stereoscopic view, commanding instrument motion through Master Tool Manipulators (MTMs). The Patient Side Manipulators (PSMs) close the loop using forward and inverse kinematics to drive actuators, generating low-level feedback (e.g., joint states, end-effector poses) that complement video. Such multimodal signals are valuable for activity recognition Huaulmé et al. [2023] and for safety/error detection Yasar and Alemzadeh [2020], particularly when video is degraded by occlusion or lens contamination Allers et al. [2016].

Although Intuitive Surgical's da Vinci systems remain the only FDA approved robots widely used in operating rooms, the ecosystem is expanding. Medtronic's Hugo platform received European approval for urology and gynecology procedures in 2021 Cepolina and Razzoli [2022], and Medicaroid Corporation's Hinotori surgical robot has been employed for actual surgeries Miyamoto et al. [2025]. Research groups routinely employ open-source and industrial robots adapted for surgical tasks, including Raven-II Hannaford et al. [2012], Taurus II, and ABB's YuMi Madapana et al. [2019]. Simulation environments are also prevalent, e.g., SimNow (Intuitive), and virtual models of Raven-II Li et al. [2016] and Taurus II Madapana et al. [2019]. A broader survey of commercial and prototype systems appears in Cepolina and Razzoli [2022]. To promote cross-platform development, the community has introduced the Collaborative Robotics Toolkit (CRTK), a common interface spanning Raven-II and the da Vinci Research Kit (dVRK) Su et al. [2020].

However, access to time-synchronized, multimodal RAS data is limited by proprietary systems, institutional privacy constraints, and restricted access to clinical platforms. Open research systems like Raven-II and dVRK mitigate some barriers Hannaford et al. [2012], Kazanzides et al. [2014], but remain hardware-dependent and require specialized access, constraining scale and heterogeneity. Furthermore, most large scale RAS datasets are vision only Psychogyios et al. [2023], Huaulmé et al. [2021], and multimodal datasets (robot kinematics and vision) are limited to dry-lab tasks Madapana et al. [2019], Gao et al. [2014]. A more recent work introduced a multimodal wet-lab dataset for imitation learning in robotic cholecystectomy, comprising the endoscopic video and internally recorded PSM and MTM kinematics data from a dVRK platform Hansen et al. [2026]. While a positive step, the resulting dataset, primarily aimed at imitation learning research, is collected using only a single unit of dVRK from short trials of clipping and cutting actions performed by two non-clinical research assistants.

The need for accessible data pipelines is growing as ML methods for training, skill assessment Fard et al. [2018], Wang and Majewicz Fey [2018], and safety monitoring Yasar and Alemzadeh [2020], Hutchinson et al. [2022] mature. Yet, there is no affordable, non-invasive, open-source framework that adapts across commercial and non-commercial robots with differing interfaces. Moreover, moving beyond exclusive reliance on robot kinematics to external sensing (e.g., hand-motion tracking Goldbraikh et al. [2022]) enables platform-agnostic acquisition, broadening participation, and improving generalizability.

To address these gaps, we present an open-source, platform-agnostic multimodal data collection system for RAS research that provides time-synchronized streams and modular expansion of sensing modalities without any impact to the robotic platform and minimal interference with user experience. In summary, our contributions are the following:

1. We present an open-source Multimodal Data Acquisition System (MiDAS) that can be non-invasively integrated with standard teleoperated robotic surgery systems to collect fully synchronized data in real-time.
2. We propose new alternative modalities to proprietary robot kinematics, including surgeon hand and foot motion kinematics (captured using electromagnetic hand tracking sensors, depth cameras, and force sensors) and demonstrate that they closely approximate surgeon and robot kinematics and can be effectively used for key tasks such as activity recognition.
3. We evaluate MiDAS by collecting data from two RMIS tasks: standard Peg Transfer on Raven-II and a novel multimodal dataset of hernia repair suturing on realistic KindHeart tissue models using the da Vinci Xi, performed during a training bootcamp at the University of Virginia Hospital. We identify effective modality combinations for multimodal gesture recognition by benchmarking state-of-the-art (SOTA) Transformer Weerasinghe et al. [2024] and CNN models Farha and Gall [2019].
4. We publicly release both the data collection system and the resulting multimodal, multi-platform datasets with gesture annotations, providing the community with a unique testbed for exploring multimodal learning and alternative sensing for activity recognition in RMIS.

2 Background and Related Work

2.1 Open-source Data Acquisition Systems for RAS

Several integrated data acquisition systems have been developed for surgical robotic platforms. Oh et al. [2024] implemented a system capturing MTM, PSM, Endoscopic Camera (ECM), and pedal states from dVRK robots through joint states, forward kinematics, optical pose estimation via fiducial markers, and hand-eye calibration. Pedal data were obtained from internal message streams and electrosurgical unit signals. They evaluated the usefulness of the collected data (from training on cholecystectomy procedures) for pedal press prediction and object segmentation. While the dataset contains multimodal data, it does not include gesture- or activity-level annotations to support surgical workflow modeling or activity recognition tasks. In addition, the reliance on platform-specific internal data streams

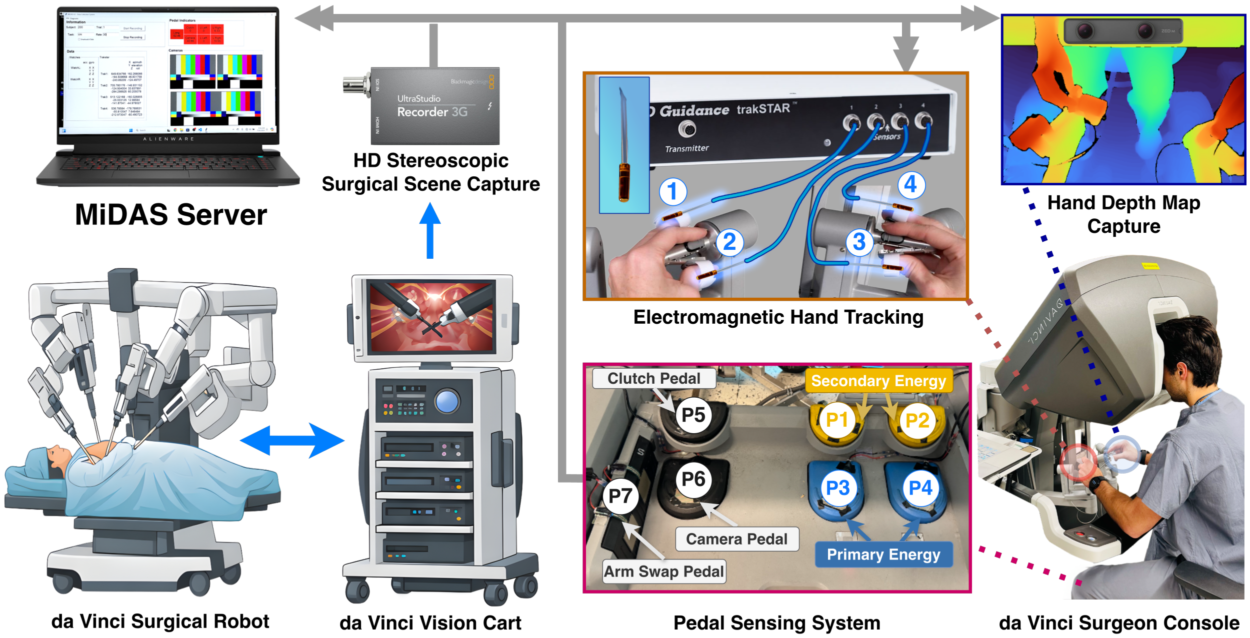


Figure 1: The Data Collection System deployed on a da Vinci Surgical Xi system.

limits generalizability to other systems and clinical settings. Hashemi et al. [2023] developed a vision-based system using depth cameras to track surgeon hand movements as MTM proxies, with surgical scene video captured from the vision tower and pedal/clutch states extracted from on-screen graphical indicators. Based on the findings by Hutchinson [2023] regarding electromagnetic (EM) tracking for surgical robot kinematics, we propose EM tracking as an effective alternative to visual modalities for capturing MTM and PSM kinematics. This work evaluates EM tracking alone and combined with other external sensing modalities that eliminate the need for internal robotic system access.

2.2 Hand Tracking for Surgical Robotics

Prior work has examined commercial hand-tracking systems for surgical contexts. Optical sensors such as Leap Motion and Kinect showed major limitations. Leap Motion had poor dynamic accuracy and range Guna et al. [2014] and both lacked the precision and robustness needed for surgical interfaces Kim et al. [2014]. More recent work has utilized EM tracking. Goldbraikh et al. [2022] applied trakSTAR EM sensors to surgeons' hands for kinematic gesture recognition during suturing, achieving high accuracy (82.4%) with models like MS-TCN++. Goldbraikh et al. [2023] extended this with new temporal architectures and data augmentation methods, evaluated on multiple dry-lab open surgery datasets (variable tissue Goldbraikh et al. [2022], basic suturing/knot-tying Gao et al. [2014]), establishing new baselines for surgical motion analysis.

MiDAS leverages EM Hand Tracking (EmHT) to non-invasively approximate end-effector kinematics when direct robot signals are inaccessible. EmHT's occlusion robustness and absolute position inference simplify deployment around consoles and patient carts. We quantify approximation fidelity by benchmarking the EM-derived trajectories against ground-truth robot kinematics on the Raven-II, enabling transparent error characterization and reproducible evaluation.

3 MiDAS: A Multimodal Data Acquisition System

We present MiDAS, a platform-agnostic, open-source, non-invasive system that records synchronized multimodal data during teleoperated RAS to approximate robot kinematic trajectories. Unlike existing setups constrained to proprietary robot telemetry, MiDAS integrates external sensing modules including EmHT, a depth hand capture camera, and foot pedal force sensors through a unified software framework. The system supports real-time synchronous acquisition across heterogeneous data streams, enabling reproducible data collection on both open research platforms (Raven-II) and commercial systems (da Vinci Xi) maintaining a platform-agnostic nature.

3.1 High-level Architecture

We employ a client–server architecture for coordinated, real-time acquisition of heterogeneous data streams. A user-friendly GUI supports metadata configuration, modality selection, and system monitoring. Full architectural and synchronization details are provided in *Appendix A.1*.

3.2 3D Electromagnetic Hand Tracking

To capture surgeon hand motion non-invasively, we employ the NDI *trakSTAR* electromagnetic tracking device Northern Digital Inc. [2024] with four miniature 6-DoF sensors (3D position and orientation). Two sensors are mounted on each MTM control at the thumb and middle-finger contact pads (Figure 1), preserving the console hardware while measuring fine manipulator motions. This placement yields continuous finger-level pose trajectories during operation.

MiDAS logs 3D position and orientation for each sensor at 270 Hz. To use these measurements as a proxy for robot kinematics (MTM/PSM *position, orientation, velocity, grasper angle*), raw sensor data are mapped from the electromagnetic tracker frame to the robot reference frames via a calibrated rigid-body transformation (rotation and translation). To account for residual modeling errors, sensor noise, and controller inaccuracies, we augment this transformation with a learned residual modeled using a Multi-Layer Perceptron (MLP). This hybrid analytical and data-driven approach enables accurate approximation of robot end-effector trajectories using only external sensing. We evaluate the accuracy and correlation of the resulting trajectories against ground-truth MTM/PSM poses on the Raven-II platform in Sec. 5.1.1, with additional details in *Appendix A.2*.

3.3 RGB-D Hand Tracking

To capture fine-grained hand movements, including manipulating finger positions within the MTM workspace, we mount an RGB-D stereo camera above the surgeon console. Specifically, we use a ZED Mini camera Stereolabs Inc., which provides synchronized 720p color video and dense depth maps at 30 Hz, enabling markerless 3D reconstruction of hand poses and manipulation trajectories without instrumenting the surgeon.

To approximate robot end-effector motions, we extract thumb and middle-finger 3D positions for each hand and transform them into the MTM frame using a lightweight MLP, yielding keypoint features denoted as *HandKP*. Limited camera field-of-view (FoV) and occlusions cause intermittent keypoint detection gaps. We fill gaps of at most one second using cubic spline interpolation and longer gaps using forward-filling. Details on keypoint detection, HandKP-MTM transformation, and camera calibration are provided in *Appendix A.3*.

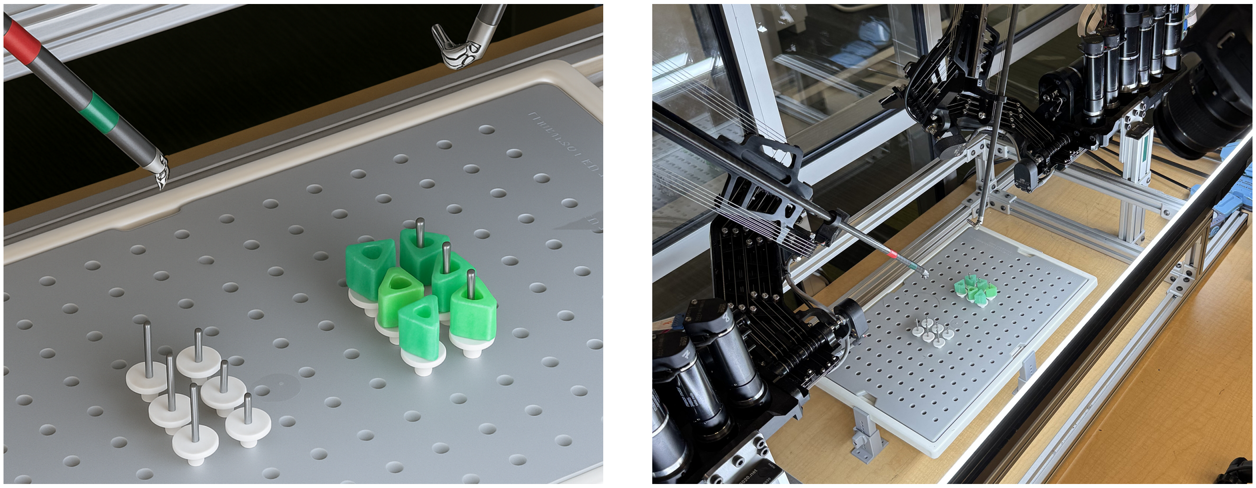
3.4 HD Stereoscopic Video Capture

We use Open Broadcaster Software (OBS) OBS Project [2025] as a unified capture layer to support stereoscopic video recording across platforms. On da Vinci Xi, stereoscopic SDI output is captured using Blackmagic SDI recorders, while on Raven-II, stereo video is recorded using a ZED camera. OBS records synchronized, high-definition stereo streams at 30 Hz. Recording is controlled programmatically via the OBS WebSocket API, enabling deterministic session control and integration with our server.

3.5 Pedal Sensing System (PSS)

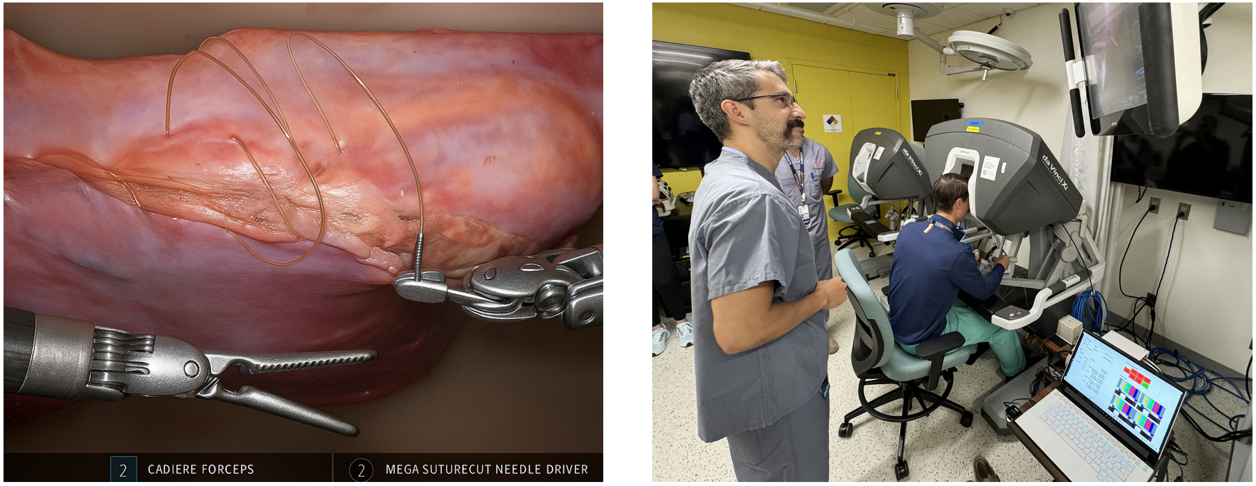
The Pedal Sensing System (PSS) captures surgeon foot pedal interactions, which encode essential behavioral cues during robot-assisted surgery including instrument activation, clutching, and potential markers of expertise. To detect pedal presses non-invasively, we developed a *compact, low-cost, open-source module* built around an Arduino-class microcontroller and thin-film force-sensitive resistors (FSRs) Industries [2012]. Sensors are affixed directly to the console pedal surfaces, enabling continuous monitoring without modifying hardware or altering the surgeon's tactile experience.

PSS supports up to nine pedals and records both binary pedal states and timestamped analog signals at 30 Hz, providing detailed activation traces for downstream analysis. During initial setup, each sensor is mapped to the corresponding robot pedal, followed by a brief calibration to determine pedal-specific thresholds for reliable actuation detection. These thresholds are further refined using a data-driven procedure based on ground-truth pedal labels. The complete hardware design, calibration steps, signal-processing details, and threshold-optimization methodology are provided in *Appendix A.4*.



(a) Raven-II Peg Transfer Task

(b) Raven-II Dry-Lab Setup



(c) da Vinci Xi Suturing Task

(d) Surgical Bootcamp

Figure 2: Experimental setups for multimodal data collection on Raven-II and da Vinci Xi systems, spanning dry-lab and clinical training environments.

4 Datasets

To demonstrate the effectiveness of our system, we collected and publicly release two multimodal datasets acquired on two RMIS platforms (Figure 2), providing high-fidelity benchmarks for future research. Uniquely, this collection introduces synchronized multimodal data recorded during realistic hernia repair procedures on KindHeart simulation models, bridging the gap between abstract dry-lab tasks and clinical reality.

4.1 Raven-II

We collected 15 Peg Transfer Ritter and Scott [2007] trials on the Raven-II platform, resulting in 36 minutes of data with gestures annotated using the DESK taxonomy Madapana et al. [2019]. Each trial involved moving three pegs across the pegboard, performed by a graduate researcher with prior experience. See Table 1 for dataset distribution and statistics. Ground-truth kinematic data, including MTM and PSM position, orientation, velocity, grasper angle, and clutch pedal presses were available from the open source platform and are used for system evaluation.

4.2 da Vinci Xi

We deployed MiDAS during a robotic surgery training bootcamp at the University of Virginia (UVA) hospital. Participants performed 7 trials of inguinal hernia repair and 10 trials of ventral hernial repair, among other procedures using

Dataset	Tasks	Data Modalities				Gesture Annotations			
		PSM	MTM	Video	Pedals	ID	Gesture Name	Samples	Avg. Dur. (s)
Raven-II *Dry-lab* (15 trials)	Peg Transfer	✓	✓	✓	✓	S1	Approach peg	49	5.84 ± 3.66
						S2	Align & grasp	50	6.05 ± 7.22
						S3	Lift peg	51	2.83 ± 1.28
						S4	Transfer peg – Get together	52	8.08 ± 4.04
						S5	Transfer peg – Exchange	50	9.65 ± 6.85
						S6	Approach pole	47	5.87 ± 2.52
						S7	Align & place	46	5.37 ± 2.74
							Total: 7	**345**	**36 (min)** ∼2.4 mins/trial
da Vinci Xi *High-fidelity* *Hybrid Porcine* *Tissue, ex vivo* (17 trials)	Suturing (Inguinal and Ventral Hernia Repair)	†	†	✓	✓	G1	Orient Needle	246	9.30 ± 9.48
						G2	Target Needle	321	3.61 ± 7.63
						G3	Push Needle Through Tissue	328	7.82 ± 10.88
						G4	Pull Needle out of Tissue	273	5.72 ± 6.67
						G5	Reach Suture	39	6.10 ± 6.56
						G6	Pull Suture	233	13.12 ± 21.98
						G7	Make C Loop	143	5.07 ± 3.74
						G8	Square Knot and Cinch	141	8.47 ± 6.88
							Total: 8	**1,724**	**212 (min)** ∼12.5 min/trial
DESK Madapana et al. [2019] *Dry-lab* (76 trials, 228 transfers)	Peg Transfer	✓	×	✓	×		**Total: 7**	1,611	∼**65 (min)** ∼0.8 min/trial
JIGSAWS Gao et al. [2014] *Dry-lab* (103* trials)	Suturing (39) Knot-Tying (28) Needle-Passing (36)	✓	✓	✓	×		**Total: 11**	1,703	∼**206 (min)** ∼2 min/trial
SAR-RARP50 Psychogyios et al. [2023] *Human, in vivo* (50 trials)	Suturing (Radical Prostatectomy)	×	×	✓	×		**Total: 8**	2,219	∼**267 (min)** ∼5.34 min/trial
ImitateCholec Hansen et al. [2026] *Porcine, ex vivo* (34** trials)	Clipping & Cutting (Cholecystectomy)	✓	✓	✓	×		**Total: 17 *****	18,343	∼**1200 (min)** ∼1.1 min/trial

Table 1: Dataset statistics for gesture recognition tasks across two platforms versus datasets from related work. *One JIGSAWS trial is a single instance of three-throw suturing, tying three knots, or passing a needle through a ring three times. **Assuming a single trial corresponds to one ex vivo tissue. ***Four distinct actions (Grasp, Apply Clip, Go Back, Cut) are repeated to clip and cut left and right tubes. †-Available from MiDAS EM and RGB-D hand tracking.

two units of da Vinci Xi systems including surgeon consoles, vision towers and patient-side robotic arms. Throughout the bootcamp, sessions were conducted under direct supervision and guidance of expert resident and attending proctors at UVA, and benefited from high quality KindHeart tissue models and standard instruments and endoscopic cameras from Intuitive Surgical.

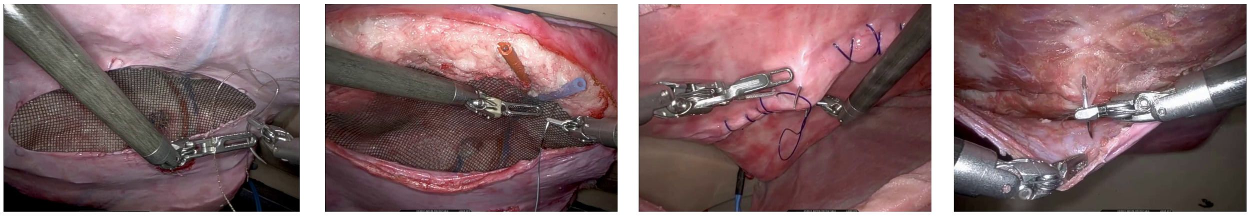
The resulting data is subsequently annotated and released for public use, aiming to promote the generation of future multimodal surgical datasets and enable development of multimodal cognitive surgical assistants by the research community. The results of a post-completion survey of the participants, and details about the taxonomy of annotations are presented next.

4.2.1 Participants Feedback Survey

A total of 40 participants, mostly 2nd and 3rd year surgical residents, performed 7 trials of inguinal hernia repair and 10 trials of ventral hernial repair, with a total of 3.5 hours of annotated suturing segments. Survey responses indicated strong acceptance of the simulation environment and supporting infrastructure, with 92.3% of participants reporting overall satisfaction with the training experience and 86.1% rating the KindHeart tissue models as realistic. Importantly, **80%** of participants **disagreed or strongly disagreed** that the sensors negatively affected their psychomotor performance, indicating that MiDAS is non-invasive and does not adversely affect participant performance.

4.2.2 Taxonomy of Suturing Gestures

We developed a structured taxonomy of robotic suturing gestures in collaboration with an expert robotic surgeon. The taxonomy builds upon established gesture definitions from prior dry-lab Gao et al. [2014] and in-vivo robotic suturing studies Psychogyios et al. [2023], and was further extended to capture nuances observed in realistic surgical workflows, including fine-grained suture thread management and tissue approximation behaviors (e.g., G6 Pull Suture). Representative examples of each gesture are illustrated in Figure 3, and detailed gesture definitions are provided in Table 2.

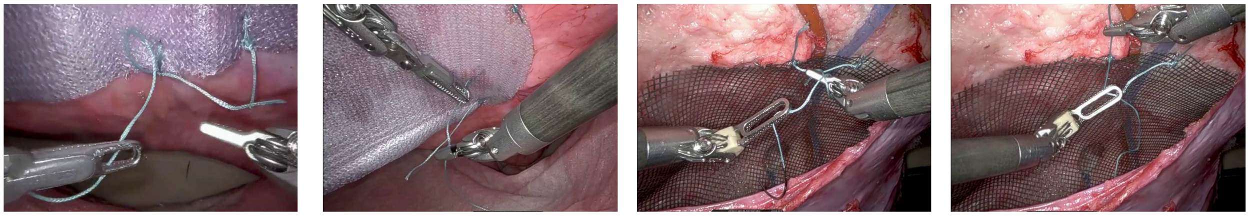


G1 - Orient Needle

G2 - Target Needle

G3 - Push Needle Through Tissue

G4 - Pull Needle out of Tissue



G5 - Reach Suture

G6 - Pull Suture

G7 - Make C Loop

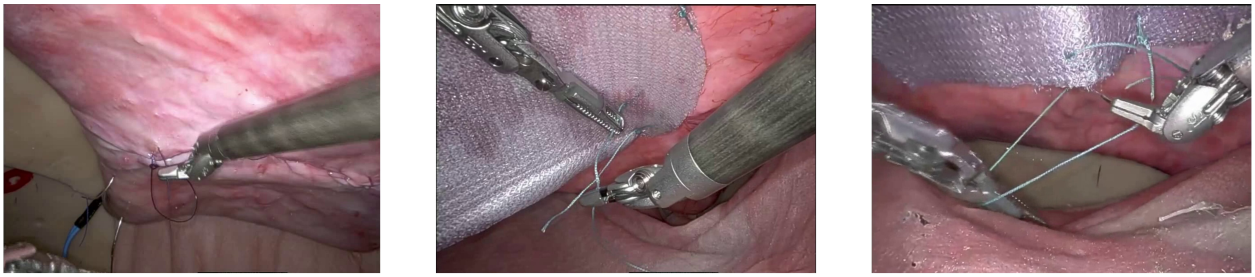
G8 - Square Knot and Cinch

Figure 3: Suturing Gestures in Inguinal and Ventral Hernia Repair with KindHeart Simulation Models.

We summarize their distribution within the MiDAS dataset and compare them against related multimodal and wet-lab suturing datasets in Table 1.

The taxonomy reflects the natural procedural progression of robotic suturing. It begins with needle orientation for proper alignment (G1), followed by needle targeting and positioning at the tissue entry point (G2). The subsequent gestures capture needle driving through tissue (G3) and needle extraction (G4). The workflow then transitions to suture management, including reaching and grasping the suture tail with the opposing instrument (G5), and performing thread manipulation to release slack, maintain workspace organization, or approximate tissue (G6). The final stages include forming the C-loop around the assisting instrument (G7) and completing square knot tying with cinching (G8). The taxonomy was iteratively refined and validated by a practicing robotic surgeon to ensure clinical relevance. A comprehensive description of gesture objectives and functional roles is presented in Table 2.

A distinguishing characteristic of our dataset is its emphasis on capturing intra-gesture variability reflective of real surgical practice. For example, suture manipulation is represented through multiple execution strategies, including single-handed pull, bimanual pull, and fulcrum-assisted pull, as illustrated in Fig. 4.



Pull Suture One Hand

Pull Suture Two Hands

Pull Suture Fulcrum

Figure 4: Suture manipulation using one hand, two hands and the fulcrum technique.

4.2.3 Data Annotation Procedure

Following our proposed taxonomy, the annotation process was conducted by a multidisciplinary team consisting of an expert surgeon, two graduate researchers, and two undergraduate researchers, working under the direct supervision of the expert surgeon, who has extensive experience in laparoscopic and robotic surgery. An open-source video annotation tool, the VGG Image Annotator (VIA) Dutta and Zisserman [2019], was customized to align with our taxonomy and to streamline the annotation workflow. Using this platform, annotators labeled the raw video data by marking temporal

Code	Gesture Name	Goal	Description
G1	Orient Needle	Needle is grasped at 50% distance from tip to end, with a 55° angle relative to the grasper.	**Main arm:** Grasps and manipulates the needle body to adjust its orientation. **Support arm:** Grasps the needle temporarily to facilitate re-orientation or hand-off.
G2	Target Needle	Needle tip contacts the tissue at the intended entry location.	**Main arm:** Positions the needle tip at the specific entry point on the tissue surface. **Support arm:** Retracts surrounding tissue to expose the target area or remains idle.
G3	Push Needle Through Tissue	Needle tip completely penetrates and exits the tissue layers.	**Main arm:** Applies force to push the needle body through the tissue. **Support arm:** Retracts tissue to stabilize the entry site or release tension.
G4	Pull Needle out of Tissue	Needle end fully exits the tissue.	**Main arm:** Grasps the emerging needle tip and pulls it clear of the tissue. **Support arm:** Counter-tracts the tissue or assists in grasping the needle if necessary.
G5	Reach Suture	Suture is grasped at a distance from the tissue that permits loop formation.	**Main arm:** Maintains position or orients itself to act as the fulcrum for the subsequent loop. **Support arm:** Grasps an appropriate point on the long end of the suture thread.
G6	Pull Suture	Suture on the exit side lengthens while the entry side shortens to close the gap.	**Main arm:** Grasps and pulls the suture thread away from the tissue. **Support arm:** Either grasps and pulls the suture (in two-handed techniques) or provides a fulcrum point for the thread.
G7	Make C Loop	A loop of suture is formed around the support instrument.	**Main arm:** Remains stationary or positions itself to act as the post for the loop formation. **Support arm:** Manipulates the long end of the suture to wrap it around the opposite jaw.
G8	Square Knot and Cinch	The knot is formed and tightened securely against the tissue.	**Main arm:** Grasps the short tail of the suture and pulls it through the loop. **Support arm:** Pulls the long end of the suture in the opposite direction to cinch the knot.

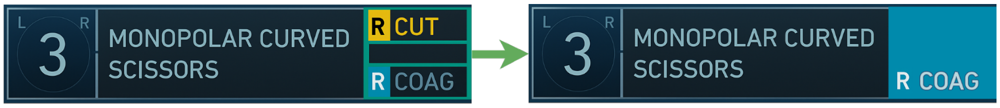
Table 2: Taxonomy of Surgical Gestures for Suturing.

gesture boundaries of when an action began and ended. These temporally segmented annotations were subsequently processed to generate frame-level gesture labels for downstream analysis.

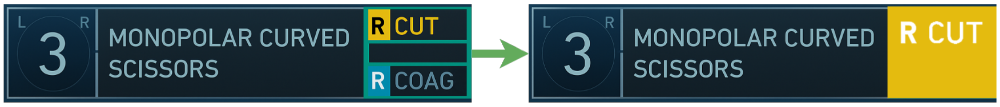
The ground-truth internal MTM and PSM kinematic data was not available from da Vinci Xi. We extracted the ground-truth pedal data from the video recordings of surgeon view by analyzing pedal state indicators (see Figure 5). Specifically, we extracted the pedal states on the user interface by analyzing certain user interface elements similar to Hashemi et al. [2023]. These elements change color or shape according to the state of their corresponding pedal. We compared the extracted pedal status data with the data from the PSS. To match PSS pedals with those of the da Vinci console, we use the mapping presented in Table 5.

4.2.4 KindHeart Surgical Models

KindHeart is a healthcare simulation company that provides real-tissue simulation in dry-lab settings using porcine tissues. Their hernia models, part of the KindHeart General Surgery Simulation system, are designed to be fitted with the KindHeart abdominal simulator and provide realistic tissue properties and anatomical fidelity that closely mirror actual surgical conditions Feins [2021]. This high-fidelity simulation environment enabled participants to practice hernia repair techniques with tissue handling characteristics comparable to live surgery, while maintaining the controlled and repeatable conditions necessary for data collection.



(a) Primary Energy Right Pedal : PSS Pedal 4



(b) Secondary Energy Right Pedal : PSS Pedal 2

(c) Clutch Control: PSS Pedal 5 **(d)** Camera Pedal : PSS Pedal 6

Figure 5: Vision-based Extraction of Ground-truth Pedal Data for da Vinci Xi.

5 Experiments

5.1 Data Validation

A key design objective of MiDAS is enabling high-fidelity, non-invasive multimodal data collection without requiring access to proprietary robot telemetry. To establish the reliability of MiDAS as a data acquisition platform, we quantitatively validate externally sensed hand and foot interaction signals against ground-truth internal kinematic and control data available from surgical robotic systems. This validation is critical for demonstrating that externally captured signals can serve as accurate proxies for operator intent and robot execution dynamics.

Specifically, we evaluate the correspondence between MiDAS sensing modalities and internal robot measurements using cross-modal temporal and trajectory alignment analyses. Hand interaction sensing modalities, including EmHT and HandKP, are compared against internal master tool manipulator (MTM) and patient-side manipulator (PSM) kinematics. Similarly, foot interaction signals captured using the Pedal Sensing System (PSS) are evaluated against ground-truth pedal activation signals. The following subsections present detailed correlation, error, and temporal alignment analyses across multiple robotic platforms and tasks.

5.1.1 EmHT and HandKP Data

We performed a cross-modal correlation analysis to validate the non-invasive EmHT and HandKP data as proxies for the ground-truth internal MTM and PSM kinematics using the data collected from Raven-II. This comparison is non-trivial given the complex control transformations (e.g., cable coupling, forward/backward kinematics) between the MTM input and the resulting PSM state. We used **Cosine Similarity** (CoS) to measure trajectory pattern alignment and **Normalized Root Mean Square Error** (NRMSE) to quantify magnitude error. Across the 15 peg transfer trials and after interpolation, HandKP features experienced a minimum of 6% and a maximum of 70% missed detections.

As summarized in Table 3, the results show strong positional correspondence for both EmHT and HandKP with ground-truth MTM and PSM. **EmHT** accurately captures the translational motion patterns of both the operator's input (MTM) and the final tool-tip (PSM) (See Figure 6), with **EmHT-MTM** and **EmHT-PSM** mean CoS exceeding 0.8 across all axes. **HandKP** also achieved robust positional similarity to MTM and PSM trajectories (See Figure 7 for MTM alignment), with **HandKP-MTM** and **HandKP-PSM** mean CoS exceeding 0.7 for all axes and superior 0.88 Z-axis alignment. We attribute the lower performance of HandKP to missed keypoint detections due to limited FoV and occlusions, and present a more in-depth analysis in *Appendix A.3*.

For orientation, EmHT captures MTM trends with reasonable fidelity, with EmHT-MTM achieving a CoS of 0.90 for Yaw, though Pitch and Roll show moderate alignment (0.61 and 0.65, respectively). Also, EmHT-PSM orientation alignment is significantly lower, with CoS values ranging from 0.32 (Roll) to 0.44 (Pitch). This discrepancy is anticipated as EmHT measures the surgeon's fingertip input, whereas PSM orientation is a computed state resulting from

Modalities	**X**		**Y**		**Z**	
	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)
EmHT-MTM	**0.81**	**17.5**	**0.86**	**24.4**	**0.88**	**16.2**
EmHT-PSM	**0.82**	**17.0**	**0.80**	**21.0**	**0.89**	**16.4**
HandKP-MTM	0.72	26.13	0.71	30.15	**0.89**	19.88
HandKP-PSM	**0.74**	**24.03**	**0.72**	**26.71**	0.88	**19.21**

Modalities	**Roll**		**Pitch**		**Yaw**	
	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)
EmHT-MTM	**0.65**	50.6	**0.61**	**23.5**	**0.90**	**24.8**
EmHT-PSM	0.32	**44.1**	0.44	38.4	0.36	36.9

Table 3: Mean CoS and NRMSE(%) of EmHT–MTM, EmHT–PSM, HandKP–MTM, and HandKP–PSM for positional (X, Y, Z) and rotational (Roll, Pitch, Yaw) trajectories over 15 peg-transfer trials.

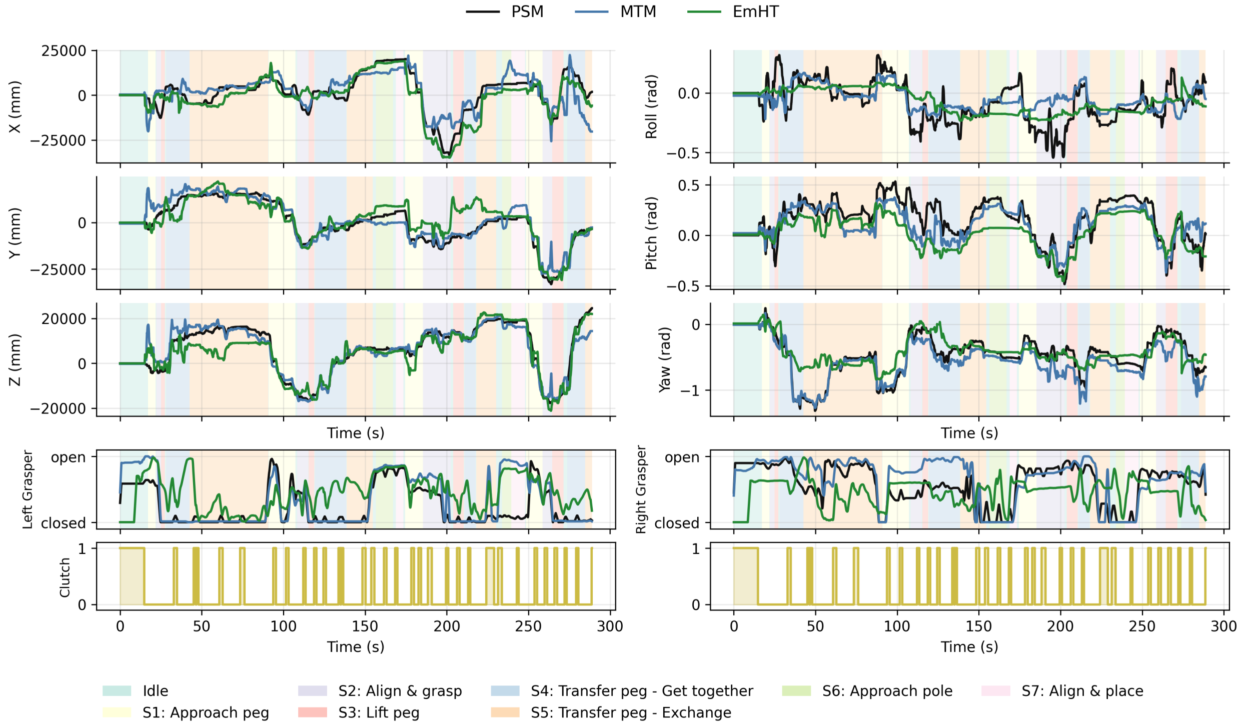


Figure 6: MTM and EmHT alignment with PSM across motion, grasper, and clutch signals for a Peg Transfer trial.

the complex control loop, causing divergence due to Raven-II's controller errors. Our method for estimating orientation from HandKP data could not achieve strong correlations with ground-truth. Results and analysis are in *Appendix A.3*.

For grasper angle, the EmHT sensors achieved moderate agreement with both MTM and PSM grasper angles (IoU $\approx$ 0.53-0.54, Accuracy $\approx$ 0.80-0.81). While EmHT can generally capture overall open/close behavior, it is not a direct replacement for capturing precise grasper angles due to noise and sensor placement strategy. More details and analysis on estimating the grasper angles from EmHT are in *Appendix A.2.2*.

5.1.2 Foot Pedal Data

We compared the binary pedal state (pressed/not pressed) trajectories collected from PSS against ground-truth pedal data (See Section 4.2) recorded from the Raven-II (1 pedal) and da Vinci systems (7 pedals). Following temporal

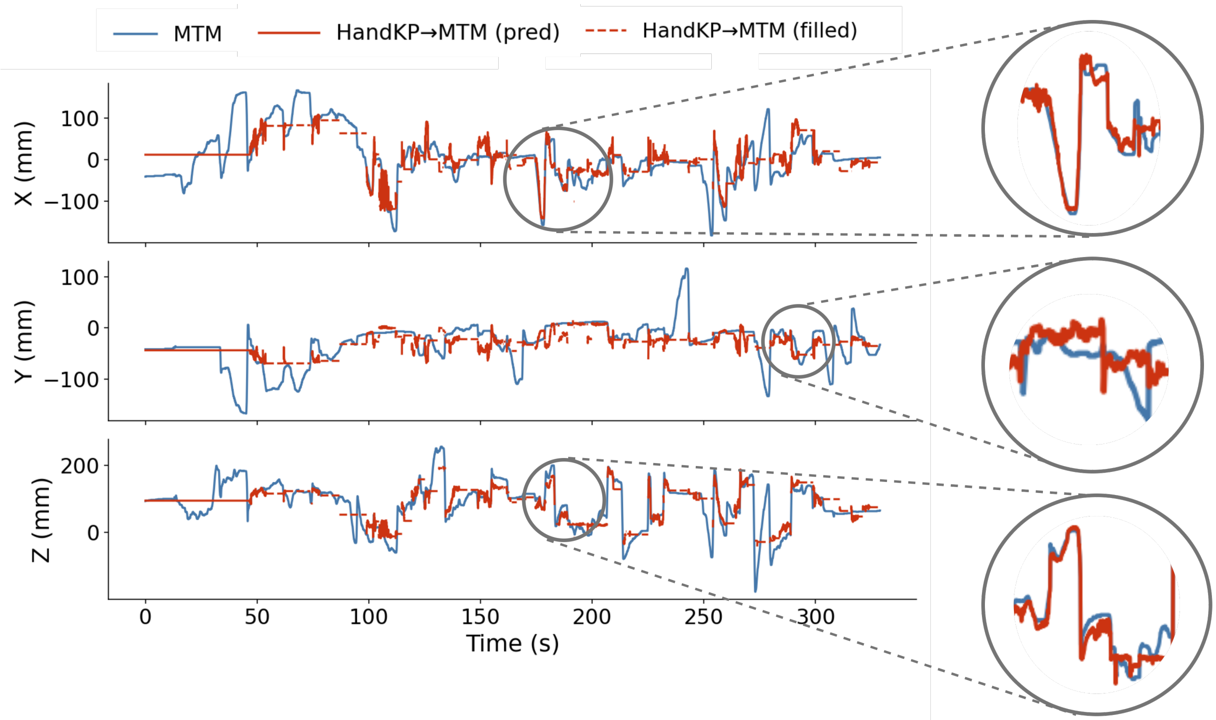
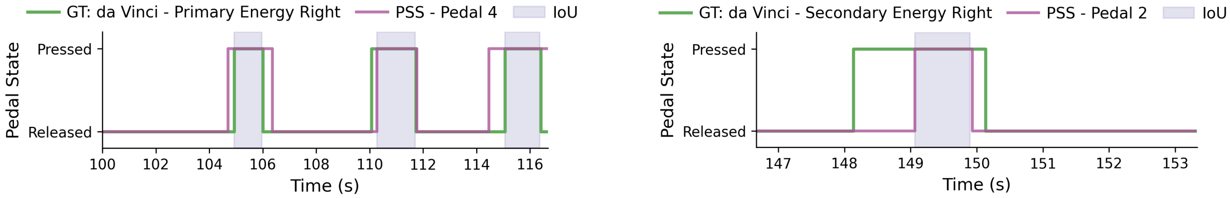


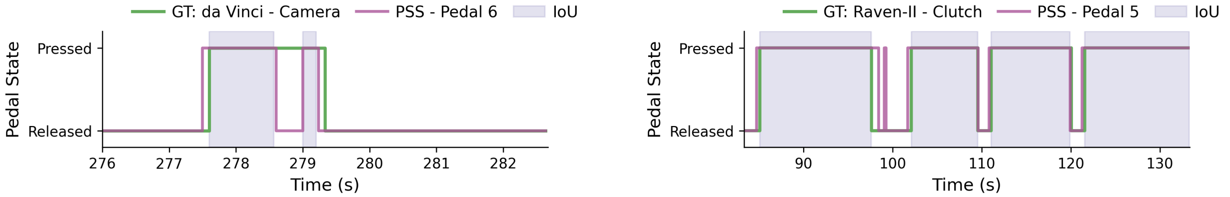
Figure 7: Estimated left hand trajectory (Left HandKP) vs. Left MTM. Average of thumb and index keypoints for each hand are transformed to MTMs frame.

synchronization, we compute frame-level precision, recall, and F1 score, along with Temporal Intersection over Union (IoU) to measure temporal overlap and residual lag to quantify system delay.



(a) da Vinci Xi Primary Energy Right GT vs PSS detection

(b) da Vinci Xi Secondary Energy Right GT vs PSS detection



(c) da Vinci Xi Camera Pedal GT vs PSS detection

(d) Raven-II Clutch Pedal GT vs PSS detection

Figure 8: Comparison of PSS pedal-state detection with ground-truth on da Vinci Xi and Raven-II data segments, with shaded IoU overlap.

Figure 8 presents a representative segment showing step plots of ground-truth and predicted pedal states for both systems. The signals exhibit strong temporal alignment during active periods, though occasional false positives are visible during transitions.

Platform	F1	Precision	Recall	IoU	Lag (ms)
Raven-II	0.85	0.78	0.95	0.74	166.67
da Vinci Xi	0.78	0.69	0.86	0.67	133.33

Table 4: Summary of pedal sensing system performance. Metrics are averaged across multiple trials.

Table 4 summarizes the quantitative results aggregated across multiple trials. For Raven-II, PSS achieves high recall (0.95), indicating that pedal presses are rarely missed, while maintaining reasonable precision (0.77), yielding an F1 score of 0.85 despite significant class imbalance favoring the unpressed state. The Temporal IoU of 0.74 indicates that approximately 74% of active pedal time is correctly captured by the system. The measured lag of 166.7 ms represents the end-to-end detection latency, which includes sensor sampling, processing, and synchronization delays. However, it should be noted that in the Raven-II platform, there is only one pedal signal (clutch) which limits the scope of this evaluation.

da Vinci pedal	PSS channel	F1	Ground-Truth Pedal Usage (%)	
			Energy Pedals	All Pedals
Secondary Energy (Left) ♦	Pedal 1	–	0.00	0.00
Secondary Energy (Right)	Pedal 2	0.83	29.21	0.34
Primary Energy (Left) ♦	Pedal 3	–	0.00	0.00
Primary Energy (Right)	Pedal 4	0.78	70.79	4.21
Clutch ★	Pedal 5	–	–	31.03
Camera	Pedal 6	0.75	–	64.40
Arm Swap ♦	Pedal 7	–	–	0.00

Table 5: Mapping between da Vinci console pedals and PSS channels with average usage over the Suturing task. Usage is reported both excluding and including the Camera pedal. ♦ Not used during Suturing. ★ Not detected using PSS, because finger clutch instead of foot pedal clutch was used.

For da Vinci Xi, the system maintains high recall (0.86) with a strong F1 score (0.77), indicating reliable detection coverage across all pedals. The lower precision (0.69) arises primarily from the intrinsic high-sensitivity design of the pedal sensing system, which is biased toward detecting all true presses (i.e., favoring recall) and therefore produces occasional false positives.

We also evaluated the per-pedal usage and detection accuracy for da Vinci Xi experiments. Pedal usage was measured using ground-truth data extracted from video and varied by procedure and console configuration. For example, *Camera* pedal was most frequently used, but the *Clutch* pedal was infrequently actuated because clutching was available on the hand manipulators. In suturing trials, considering energy pedals only, *Primary Energy (Right)* mapped to PSS *Pedal 4* accounted for roughly 70.8% of all actuations with a detection F1 of 0.78.

5.2 Downstream Evaluation for Gesture Recognition

Following the correlational analysis, we evaluate whether external modalities collected by MiDAS (EmHT and HandKP) can be used as alternatives to internal robot kinematics for downstream tasks such as surgical gesture recognition.

Our goal is not to set a new SOTA, but to examine how using individual external modalities or their fusion can impact downstream recognition performance.

We benchmark two SOTA temporal models, MTRSAP Weerasinghe et al. [2024] and MS-TCN++ Farha and Gall [2019], on MiDAS Peg Transfer (Raven-II) and Suturing (da Vinci Xi) datasets. Models are trained using single EmHT and HandKP modalities and their fusion with image features (EmHT-Image, HandKP-Image), and compared against ground-truth kinematic (MTM/PSM on Raven-II) and image-only baselines. Proprietary PSM/MTM kinematics are unavailable on da Vinci Xi; thus, image features serve as the only baseline. Raven-II data is downsampled from 30 Hz to 10 Hz, while da Vinci Xi data is used at its native 30 Hz. All models use identical architectures, training protocols, and train/validation/test splits. To assess baseline validity and the effect of dataset size, we additionally evaluate kinematic- and image-only models trained on larger related datasets: DESK Madapana et al. [2019] (processed via COMPASS Hutchinson et al. [2023]) for Peg Transfer, and SAR-RARP50 Psychogyios et al. [2023] and SAIS Kiyasseh et al. [2023] for Suturing.

We also performed a comprehensive ablation study on the effect of using different kinematic features such as velocity, grasper angle, and clutch states, two different image encoders, ResNet-50 Wang and Majewicz Fey [2018] and DINOv2 Oquab et al. [2023], and domain-adaptive pretraining on larger data (e.g., DESK Madapana et al. [2019]). More details are presented in *Appendix B*.

Dataset	Model	Modality	Features	Acc	F1	Prec.	Rec.
Raven-II Peg Transfer (15 trials)	**MTRSAP** Weerasinghe et al. [2024]	**PSM**	**pos., ori., vel., grasper.**	**0.88**	**0.87**	**0.88**	**0.87**
		MTM	**pos., ori., vel., grasper.**	0.87	0.87	0.90	0.86
		Image†	RGB (DINOv2)	0.48	0.48	0.48	0.48
		EmHT	**pos., ori., vel.**	**0.87**	**0.86**	**0.87**	**0.86**
		EmHT,Image	pos., ori., vel., RGB (DINOv2)	0.58	0.56	0.60	0.58
		HandKP	pos., vel.	0.40	0.38	0.41	0.39
		HandKP, Image	pos., vel., RGB (DINOv2)	0.38	0.33	0.36	0.37
	MS-TCN++ Farha and Gall [2019]	**PSM**	**pos., ori., vel., grasper.**	0.85	0.84	0.86	0.86
		MTM	**pos., ori., vel., grasper.**	**0.86**	**0.86**	**0.86**	**0.87**
		Image†	RGB (DINOv2)	0.40	0.47	0.41	0.44
		EmHT	**pos., ori., vel.**	**0.79**	**0.78**	**0.80**	**0.81**
		EmHT,Image	pos., ori., vel., RGB (DINOv2)	0.65	0.64	0.67	0.67
		HandKP	pos., vel.	0.33	0.29	0.27	0.37
		HandKP, Image	pos., vel., RGB (DINOv2)	0.49	0.51	0.49	0.46
DESK Madapana et al. [2019] Peg Transfer (48 trials)	***MTRSAP** Weerasinghe et al. [2024]	PSM	pos., vel., ori., grasper.	0.91	0.91	0.92	0.92
		Image	RGB (DINOv2)	0.78	0.78	0.77	0.78
	***MS-TCN++** Farha and Gall [2019]	PSM	pos., vel., ori., grasper.	0.92	0.93	0.92	0.92
		Image	RGB (DINOv2)	0.74	0.78	0.76	0.77

Table 6: Peg Transfer on Raven-II Gesture Recognition using Different Modalities. Best Performing Modality/Feature is reported. † Image-only modalities underperform due to limited data. * Same baselines evaluated on a larger dataset for comparison. *Abbreviations: pos.=position; vel.=velocity; ori.=orientation; grasper.=grasper angle.*

Table 6 reports the effects of different modalities on gesture recognition in Peg Transfer. As expected, MTM/PSM kinematics form strong baselines, reflecting the fidelity of robot telemetry for modeling task dynamics. Image-only models underperform compared to SOTA due to limited data (15 vs. 48 trials). Notably, using non-invasive **EmHT** achieves comparable performance to MTM/PSM, with MTRSAP model gaining an F1-score of 0.86 and accuracy of 0.87, closely matching baseline MTM/PSM models. This demonstrates that EmHT can effectively be used as an alternative to robot kinematics for gesture recognition. Using **HandKP** as the alternative modality for gesture recognition leads to suboptimal performance compared to MTM/PSM (F1-score of 0.4 vs. 0.87-0.88). This is due to suboptimal pose approximation by HandKP and missed hand detections (as discussed in Section 5.1.1).

Dataset	Model	Modality	Features	Acc	F1	Prec.	Rec.
da Vinci Xi Suturing (17 trials)	**MTRSAP** Weerasinghe et al. [2024]	Image	RGB (DINOv2)	0.69	0.67	0.65	0.65
		EmHT	pos., ori., vel.	0.64	0.60	0.63	0.61
		EmHT, Image	**pos., ori., vel., RGB**	**0.71**	**0.70**	**0.70**	**0.69**
		HandKP	pos., vel.	0.40	0.35	0.37	0.36
		HandKP, Image	pos., vel., RGB	0.58	0.58	0.58	0.57
	MS-TCN++ Farha and Gall [2019]	Image	RGB (DINOv2)	0.57	0.57	0.53	0.51
		EmHT	pos., ori., vel.	0.57	0.50	0.50	0.53
		EmHT, Image	**pos., ori., vel., RGB**	**0.63**	**0.56**	**0.64**	**0.63**
		HandKP	pos., vel.	0.43	0.36	0.39	0.37
		HandKP, Image	pos., vel., RGB	0.58	0.54	0.58	0.54
SAR-RARP50 Psychogyios et al. [2023] Suturing (50 trials)	**ASFormer Yi et al. [2021]**	Image	RGB (I3D Carreira and Zisserman [2017])	0.82†	–	–	–
	MViT Fan et al. [2021]	Image	RGB (ViTDosovitskiy [2020])	0.79†	–	–	–
SAIS Kiyasseh et al. [2023] Suturing (78 trials)	**Vision Transformer** Kiyasseh et al. [2023]	Image	RGB	–	0.50†	–	–

Table 7: Suturing on the da Vinci Xi Gesture Recognition using Different Modalities. Best results per model in **bold**. *Abbreviations: pos.=position; vel.=velocity; ori.=orientation; †=As reported or estimated from published results.*

Table 7 shows the gesture recognition results from realistic trials of Suturing on the da Vinci Xi platform. **EmHT** performs comparably to vision-only model (RGB) and further improves accuracy when fused with them, with MTRSAP **EmHT+Image (DINOv2)** achieving Acc of 0.71 and F1-score of 0.70, outperforming each single modality. For MS-TCN++, fusion helps mainly with DINOv2 (Acc 0.63, F1 0.56). These results support non-invasive tracking as

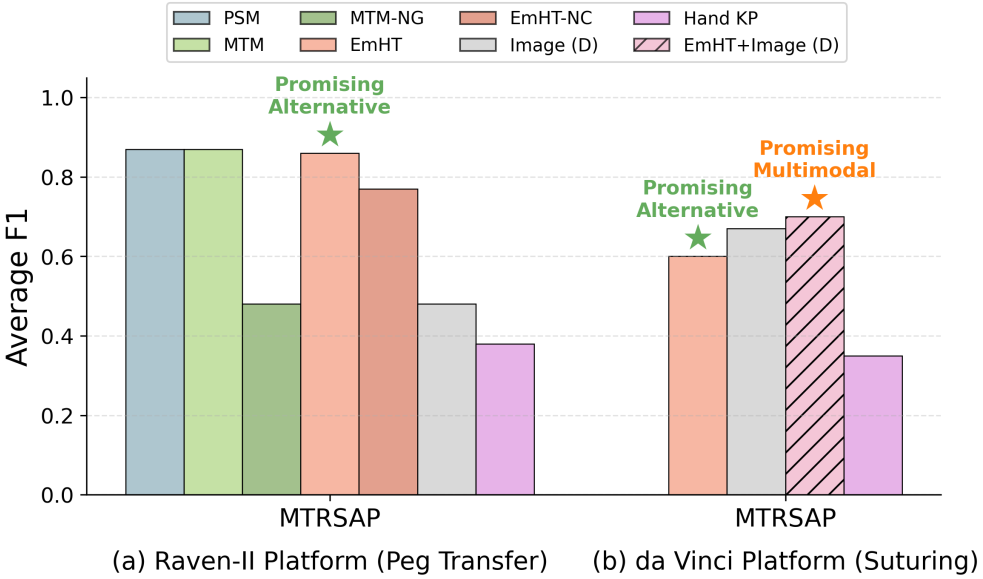


Figure 9: Best model accuracy across selected modalities on both platforms. We could not compare against PSM on da Vinci due to lack of access during data collection. *Abbreviations: NG–No Grasper Angle; NC–No Clutch State; (D)–DinoV2.*

an effective alternative and complement to vision. Similarly, (**HandKP-Image**) fusion leads to some improvement compared to image-only MS-TCN++ baseline.

Overall, in low-data settings, low-dimensional motion signals remain reliable, and vision adds value primarily when fused with complementary cues (e.g., MTRSAP). Figure 9 highlights the key modalities in both evaluations. Consistent with prior work, image-only baselines underperform with limited data, whereas robot kinematics and EmHT are robust; vision models typically require larger datasets or careful feature extraction/fusion. Thus, non-visual modalities are practical surrogates when proprietary telemetry is unavailable or video quality varies Weerasinghe et al. [2024], Zappella et al. [2013], Shi et al. [2022], Ahmidi et al. [2017], Goldbraikh et al. [2022].

6 Conclusion

In this paper, we introduced **MiDAS**, a platform-agnostic, non-invasive multimodal data acquisition system for robot-assisted surgery. MiDAS enables synchronized collection of EmHT, optical RGB-D keypoints, and pedal interaction signals without access to proprietary robot hardware or software. Correlation analysis against Raven-II ground-truth kinematics shows that electromagnetic tracking reliably approximates robot end-effector motion, achieving sub-20% normalized RMSE and high cosine similarity. Downstream gesture recognition experiments further demonstrate that external sensing modalities, particularly EmHT, achieve performance *competitive* with proprietary robot kinematics and outperform vision-based approaches due to their robustness to occlusions and out-of-view conditions. These results highlight non-invasive motion sensing as a practical surrogate for internal robot telemetry in real surgical environments. By releasing the MiDAS system, multimodal datasets, and baseline benchmarks, we aim to support reproducible, cross-platform research, reduce reliance on closed robotic systems, and accelerate progress in multimodal learning for surgical activity understanding.

Acknowledgments

This work was partially supported by research awards from National Science Foundation (CNS-2146295) and the Commonwealth Cyber Initiative. We thank Kay Hutchinson, Shrisha Yapalparvi, and Sara Liu for contributions to the data collection system and early drafts; Tessa Heick for assistance with data collection; and Alex Farmer and Maximilian Meer for data annotation. Special thanks to Christopher Scott, MD (now at Mayo Clinic Jacksonville), Shane Strike and Mike Lewis (Intuitive Surgical), and the University of Virginia Department of Surgery for their help with organizing the Bootcamp.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability

The multimodal datasets and data collection system introduced in this study are publicly available at `https://uva-dsa.github.io/MiDAS/`.

Ethics Approval and Informed Consent

The surgical bootcamp and associated data collection were approved by the University of Virginia Institutional Review Board under protocol IRB-SBS 6724. All participants provided informed consent prior to participation.

Technical Appendix

MiDAS: A Multimodal Data Acquisition System and Dataset
for Robot-Assisted Minimally Invasive Surgery

Table of Contents

A Data Collection System & Correlation Analysis

A.1 High-level Architecture

We adopt a client-server architecture for data acquisition. A central server coordinates multiple client processes that stream heterogeneous data (e.g., kinematics, video) as shown in Figure 10 at different rates, implemented via Python `multiprocessing` workers to enable concurrent, real-time ingestion and buffering. The server provides a GUI (Figure 11) to (i) enter study metadata (subject, procedure, trial) and master sampling frequency f; (ii) select modalities to be acquired at the target rate f; and (iii) monitor live system status of all connected clients through a health dashboard. The modular code, implemented in Python, separates transport, scheduling, and I/O components, allowing easy extension to new modalities and deployment scenarios.

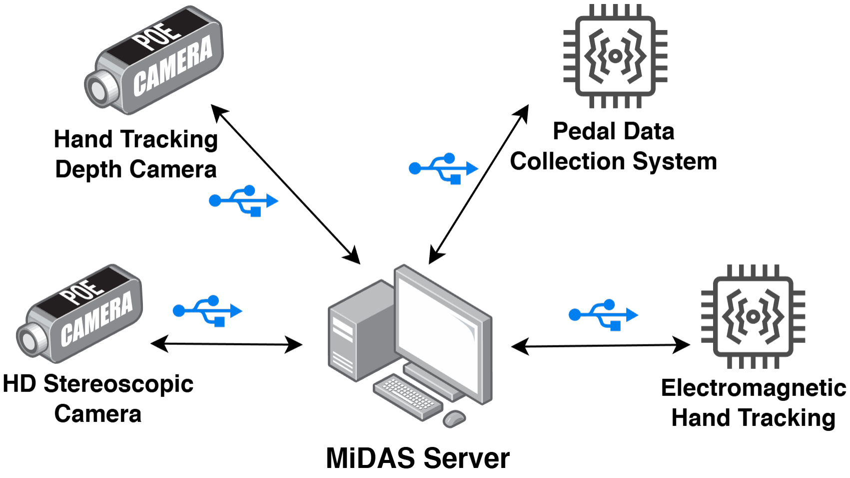
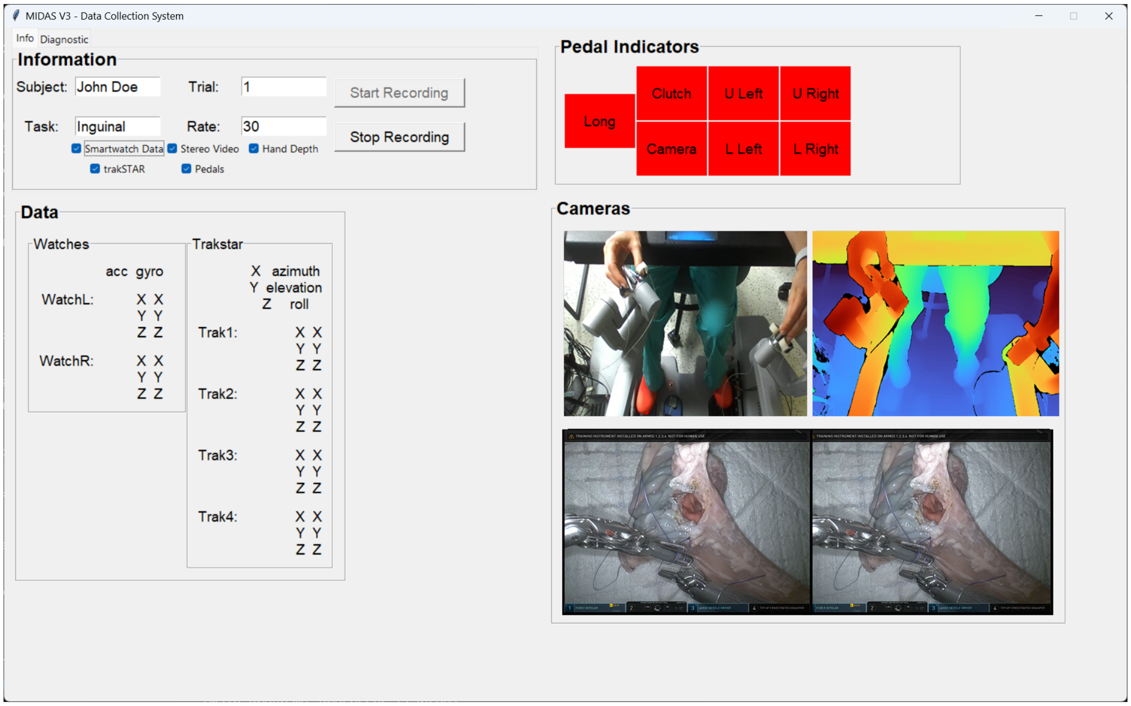


Figure 10: High-level architecture of MiDAS.



MIDAS V3 - Data Collection System

Subject: John Doe

Trial: 1

Task: Inguinal

Rate: 30

☑ Smartwatch Data ☑ Stereo Video ☑ Hand Depth

☑ trakSTAR ☑ Pedals

Figure 11: GUI of MiDAS Server. The procedure information such as subject, trial, task and acquisition rate are entered by user. The various pedal states, values of EmHT sensors, RGB-D view of hands and stereo surgical view are displayed to the user.

A.1.1 Synchronization

All streams are stamped on arrival at the MiDAS server with a Unix epoch timestamp, providing a unified, server-side time base that is insensitive to client clock drift. During post-processing, we treat the OBS video frame timestamps as the reference clock and align all other modalities to this time base via their server timestamps (causal frame-wise association to the nearest video time). We then visualize the synchronized signals (video, kinematics, and pedal sensor traces) for manual spot-checks to confirm alignment quality at frame granularity.

A.1.2 Estimated Cost of MiDAS

A key design goal of MiDAS was to minimize system cost in order to encourage broad adoption by the research community. At the time of development and experimentation (Q1 2024), the estimated cost for a single surgeon-console MiDAS setup was approximately \$8,000 (USD). This cost includes an NDI *trakSTAR* 3D Guidance electronics unit and accessories (∼\$4,850), Blackmagic SDI recorders and cables for stereoscopic video capture (2×, ∼\$400 total), a ZED Mini RGB-D camera for hand-depth capture (∼\$400), a custom pedal sensing system (∼\$150), and an appropriate laptop workstation to host the MiDAS server (∼\$2,000). Overall, the complete hardware setup totals approximately \$8,000 at the time of development.

A.1.3 Data Recording Usage Rate

When recording at a master synchronization frequency of 30Hz, MiDAS captures stereoscopic surgical scene video at 1080p resolution, electromagnetic hand tracking data, RGB-D hand depth data, and pedal sensing signals, resulting in an aggregate storage requirement of approximately 0.4GB/minute. This data rate can be reduced by lowering the recording frequency and/or decreasing the resolution of the surgical scene video, depending on the requirements of the target application.

Modality	Sensing Device	Average Latency (ms)
Electromagnetic Hand Tracking	NDI *trakSTAR* Northern Digital Inc. [2024]	43.3
Hand Depth	ZED Mini Stereo Camera Stereolabs Inc.	24.66
Surgical Scene Video†	Blackmagic SDI Recorder (OBS)	N/A
Foot Pedal States	Pedal Sensing System (PSS)	12.97

Table 8: Average data recording latency across MiDAS sensing modalities. Latency is computed using source-to-disk timestamps for modalities that expose device-side timestamps. †Surgical scene video captured via SDI and recorded through OBS does not expose source-side timestamps, and therefore comparable end-to-end recording latency cannot be reliably estimated under the timestamp-based definition.

A.1.4 Data Recording Latency

Data recording latency is quantified as the average time delay between sensor data transmission from the acquisition device to the MiDAS server and the subsequent persistent storage operation. For example, electromagnetic hand tracking data acquired by the *trakSTAR* 3D Guidance electronics unit are transmitted to the MiDAS server via USB serial communication. The acquisition unit assigns a transmission timestamp to each data packet, and the MiDAS server appends a secondary timestamp upon committing the data to disk.

This source–destination timestamp–based mechanism is employed for all sensing modalities where device-side (source) timestamps are available to facilitate latency characterization. For modalities that do not expose source-side timestamps, such as surgical scene video captured via SDI output from the da Vinci vision cart and recorded through OBS, only destination side timestamps at write time are available, and thus comparable end-to-end recording latency cannot be reliably estimated under this definition.

Studies indicate surgeons have average baseline reaction times (318–330 ms) that slow significantly under cognitive load Pfister et al. [2014]. During active robotic manipulation, dual-task experiments show reaction times extending to 0.4–0.9 s, increasing with complexity Lim et al. [2023], Yang et al. [2024]. Thus, the practical operative response window (400–900 ms) is substantially slower than the isolated 300 ms baseline. Based on the mean recording latencies for each sensing modality reported in Table 8 and processing times that are within few milliseconds scale, we propose the feasibility of MiDAS for real-time downstream applications, including real-time online gesture recognition, real-time error detection, and other time-critical processing tasks.

Note that reported latencies may depend on the hardware configuration of the MiDAS host system. All measurements in this study were obtained using an Alienware m15 R7 laptop equipped with a 12th Generation Intel® Core™ i7-12700H processor, NVIDIA® GeForce™ RTX 3070 Ti GPU (8 GB GDDR6), 32 GB DDR5 memory, and a 2 TB PCIe Gen4 NVMe solid-state drive.

A.2 3D Electromagnetic Hand Tracking

To capture surgeon hand motion non-invasively, we employ the NDI *trakSTAR* electromagnetic tracking device Northern Digital Inc. [2024] with four miniature 6-DoF sensors (3D position and orientation: azimuth, elevation, roll). Two sensors are mounted on each MTM control at the thumb and middle-finger contact pads, preserving the console hardware while measuring fine manipulator motion. Sensors are indexed as: (1) left middle finger, (2) left thumb, (3) right thumb, (4) right middle finger. This placement reflects typical console usage (thumb/middle grip; index finger on the clutch) and yields continuous, finger-level pose traces during operation.

According to the manufacturer, the nominal translation range of the system is ± 76 cm in any direction, with an angular range of ± 180° in azimuth and roll and ± 90° in elevation under ideal conditions. When the field generator is mounted near the MTM console at the resting height of the joysticks, this range is sufficient to encompass the typical console hand workspace. However, the effective tracking volume and accuracy depend on local electromagnetic distortion from nearby metallic and electronic components, so the final placement should be validated experimentally in the target operating room layout. The trakSTAR system is controlled via the vendor C++ API, integrated into the MiDAS acquisition pipeline.

A.2.1 EmHT for Pose Estimation

MiDAS logs the 6-DoF pose for each sensor ($i = 1...4$) at the 270 Hz sampling rate, expressed in the EmHT's coordinate frame (T). This pose, $\mathbf{P}_{T,i}$, combines a 3D position vector $\mathbf{p}_{T,i}$ and a 3D orientation vector $\boldsymbol{\theta}_{T,i}$:

$$\mathbf{p}_{T,i} = [x_i, y_i, z_i]^T \in \mathbb{R}^3 \tag{1}$$

$$\boldsymbol{\theta}_{T,i} = [\alpha_i, \epsilon_i, \rho_i]^T \in \mathbb{R}^3 \tag{2}$$

where $(\alpha_i, \epsilon_i, \rho_i)$ represent azimuth, elevation, and roll, respectively.

These externally sensed poses can be mapped to the coordinate frames of the Master Tool Manipulator (MTM), denoted M, and the Patient Side Manipulator (PSM), denoted P. This mapping is primarily achieved via a rigid body transformation, $\mathbf{T}$, which is composed of a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. This transformation can be obtained or estimated for any robot through a calibration procedure. For Raven-II robot, we obtain the transformations from the open source MTM and control software code bases. To compensate for modeling uncertainties and sensor/controller inaccuracies, we augment this with a learned residual term, $f(\mathbf{p}; \boldsymbol{\phi})$, modeled as a simple Multi-layer Perceptron (MLP) network, with $\boldsymbol{\phi}$ representing the learnable parameters of the model.

For example, the position of a sensor $\mathbf{p}_{T,i}$ can be transformed into the MTM coordinate frame (M) to find the corresponding position $\mathbf{p}_{M,i}$ using the rigid transformation and a residual model f_M with parameters $\boldsymbol{\phi}_M$:

$$\mathbf{p}_{M,i} = \mathbf{R}_T^M \, \mathbf{p}_{T,i} + \mathbf{t}_T^M + f_M(\mathbf{p}_{T,i}; \boldsymbol{\phi}_M) \tag{3}$$

Similarly, a transformation maps the sensor's position into the PSM coordinate frame (P) using $\mathbf{T}_T^P$ (composed of $\mathbf{R}_T^P$ and $\mathbf{t}_T^P$) and a residual f_P with parameters $\boldsymbol{\phi}_P$:

$$\mathbf{p}_{P,i} = \mathbf{R}_T^P \, \mathbf{p}_{T,i} + \mathbf{t}_T^P + f_P(\mathbf{p}_{T,i}; \boldsymbol{\phi}_P) \tag{4}$$

Both MLP models f_P and f_M have 2 hidden layers with 16 neurons and use Rectified Linear Units as activation function. The models take a pose in EmHT coordinates, and transform it onto the MTM or PSM coordinates. The training objective function is root mean squared error with L2 weight decay and 50% dropout to prevent over-fitting. Training is performed using cross-validation, where each peg transfer trial is taken to be one fold of the validation, and the average metrics across folds are reported. A summary of model and training hyperparameters are presented in Table 9.

Parameter	**Value**
Number of Hidden Layers	2
Neurons per Layer	16
Activation Function	ReLU
Dropout Probability	0.5
Objective Function	RMSE
Regularization	L_2 Weight Decay
Weight Decay Coeff.	1.5
Optimizer	Adam
Learning Rate	0.005
Epochs	50

Table 9: Hyperparameters and training configuration for the residual MLP model.

A similar composition of rotations is used to transform the orientation $\boldsymbol{\theta}_{T,i}$ into the target frames.

A.2.2 EmHT for Grasper Angle Estimation

In addition to the trajectory analysis, we also conducted an analysis to assess whether the EmHT sensors mounted on the MTM can reliably estimate the actual grasper open/close states of both the MTM and the PSM. To determine whether each grasper was open or closed, we first computed the Euclidean distance between the two EmHT sensors on the MTM grasper over time. A Moving Average filter was then applied to smooth the raw signal and highlight the underlying trend. Lastly, an appropriate threshold was selected to convert the filtered EmHT signal into a binarized open/close state, as illustrated in Figure 12.

According to Table 10, the MTM and PSM grasper angles showed strong consistency (IoU: 0.90, Accuracy: 0.97, Precision: 0.97), which means that the PSM grasper accurately executes the MTM commands. However, the EmHT sensors achieved moderate agreement with both MTM and PSM grasper angles (IoU $\approx$ 0.53–0.54, Accuracy $\approx$ 0.80–0.81), indicating that while EmHT can generally capture overall open/close behavior, it is not a direct replacement for capturing precise grasper angles due to signal noise and sensor placement strategy.

Table 10: Mean Intersection over Union (IoU), Accuracy, and Precision metrics for grasper angle estimation across the PSM–MTM, EmHT–MTM, and EmHT–PSM modalities, computed over 15 left-grasper peg-transfer trials.

Modalities	**IoU (↑)**	**Acc (↑)**	**Prec (↑)**
PSM–MTM	0.90	0.97	0.97
EmHT–MTM	**0.54**	**0.81**	**0.67**
EmHT–PSM	**0.53**	**0.80**	**0.68**

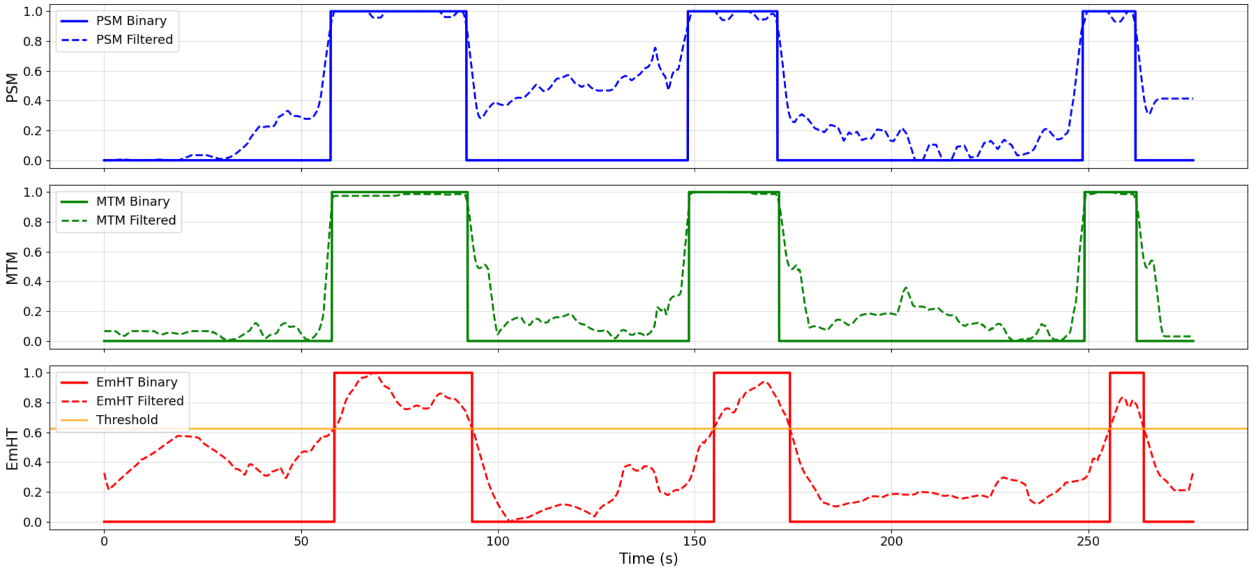


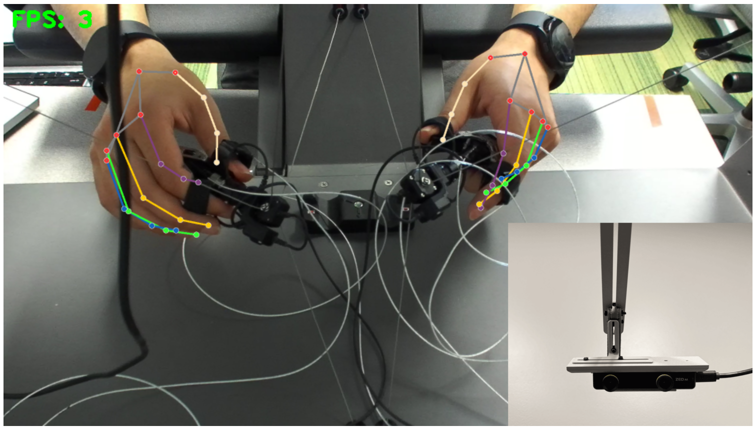
Figure 12: Left grasper angle comparison among the PSM, MTM, and EmHT measurements. Solid lines represent binarized open (0) /closed (1) states; dashed lines show normalized filtered raw signals; orange line marks the EmHT closure threshold.

A.3 RGB-D Hand Tracking

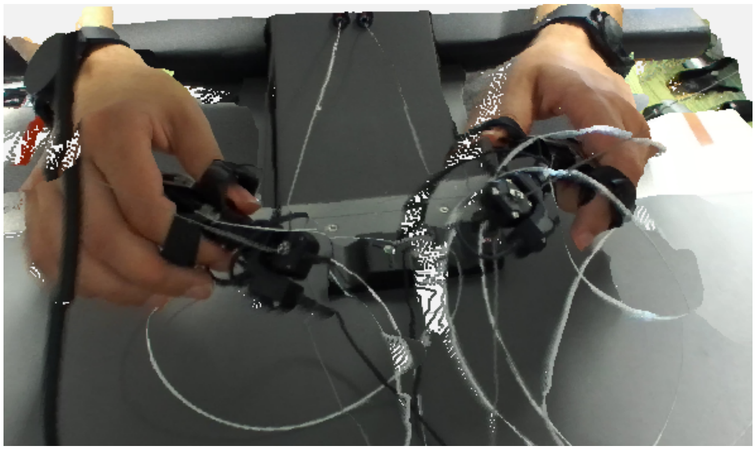
A.3.1 Keypoint Estimation and Transformation

The hand and MTM pose estimation is performed using a two-step process described below:

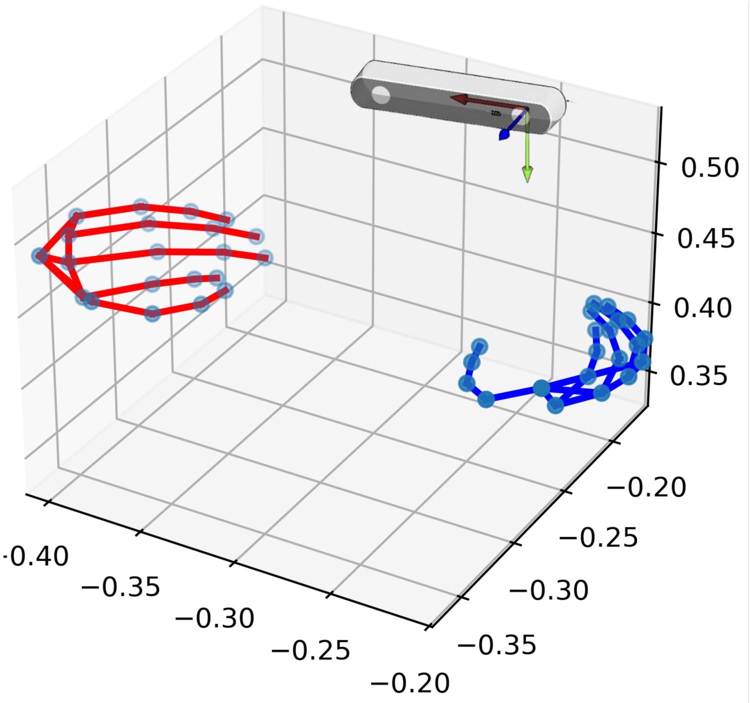
1. **Camera Calibration:** First, hand-eye calibration Horaud and Dornaika [1995] is performed between the MTMs (coordinate frame M) and the RGB-D camera (coordinate frame C) to obtain the rigid body transformation $\mathbf{T}_C^M$. This transformation, composed of a rotation matrix $\mathbf{R}_C^M$ and a translation vector $\mathbf{t}_C^M$, allows for representing all poses in the same MTM base frame. Calibration procedure is performed once per data collection session.
2. **Keypoint Extraction and Transformation:** Hand keypoints are extracted from a camera view that is aligned with a calibrated depth map, using the Mediapipe's hand landmark detection library Lugaresi et al. [2019]. An example of such keypoints is shown in Figure 13 (a). Utilizing the aligned RGB and depth frames provided by the ZED software development kit, a point cloud is generated (see Figure 13 (b)), and each extracted hand keypoint j is localized as a 3D position vector $\mathbf{p}_{C,j}$ in the camera's coordinate frame (C). Using the rigid body transformation from calibration, these keypoints are then transformed into the MTMs base frame (M) to obtain $\mathbf{p}_{M,j}$:
$$\mathbf{p}_{M,j} = \mathbf{R}_C^M\, \mathbf{p}_{C,j} + \mathbf{t}_C^M \tag{5}$$
The hand keypoints, represented in MTMs coordinate frame, are shown in Figure 13 (c). Finally, the estimated position of each MTM is taken to be the average positions of the thumb and index finger tips holding that MTM, and is referred to as HandKP.



(a) The MTMs Workspace, Hand Keypoints and the Mounted RGB-D Camera.



(b) The Point Cloud of Hands and MTMs.



(c) The localized cartesian poses of hand keypoints in MTM workspace.

Figure 13: The surgeon hands and MTM workspace, captured by the RGB-D camera and processed extract (a) hand keypoints, (b) hands and MTMs point cloud and (c) localized hand keypoints in the workspace.

As an end-to-end alternative, we also experiment with a direct transformation from hand keypoints in camera frame to MTMs base coordinates without the calibration process using an MLP network (similar to the model used in Section A.2) to assess the usability of the system for cameras with unknown calibration values. The training and evaluation also follows the same cross-validation scheme.

A.3.2 Evaluation of Accuracy

We evaluated the *HandKP* pipeline's accuracy in capturing MTM trajectories by cross-validating Rigid Transformation and MLP models over Peg Transfer trials. As detailed in Table 11, the MLP model outperformed the rigid approach, achieving robust shape similarity with mean CoS values of **0.71** to **0.97** for different axes, and NRMSEs as low as 19.8%. The primary reason for not achieving higher correlation values is the degraded performance due to occlusion and frames where the hands leave the camera field of view.

Table 11: Mean cosine similarity and NRMSE(%) RGB-D estimation of left MTM trajectory (X, Y, Z) using cross-validation over 15 Peg Transfer trials.

Model	**X**		**Y**		**Z**	
	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)
Rigid Transformation	0.45	26.39	0.63	32.14	0.79	33.06
MLP	0.72	26.13	0.71	30.15	0.89	19.88

To isolate the effect of limited camera FoV, we also evaluated either transformation method only using frames with successful hand detection and keypoint extraction, and present the results in Table 12. Using MLP-based transformations, we observe significantly higher alignment between the estimated HandKP values and actual MTM positions, with **0.91** and **0.97** CoS for X and Z axes, respectively. The magnitude errors reflected in NRMSE are also significantly lower at around 11% for X axis and a maximum of around 15%. These results establish depth-based estimation of MTM kinematics as another robust alternative to internal kinematics, provided that cameras with larger FoVs are used. We utilized trajectories transformed through the MLP method in the final accuracy evaluations and gesture recognition experiments.

Table 12: Mean cosine similarity and NRMSE(%) RGB-D estimation of left MTM trajectory (X, Y, Z) using cross-validation over 15 Peg Transfer trials (Ignoring Missing Data).

Model	**X**		**Y**		**Z**	
	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)	CoS (↑)	NRMSE (↓)
Rigid Transformation	0.81	16.48	0.67	14.75	0.93	16.00
MLP	**0.91**	**10.96**	**0.72**	**14.19**	**0.97**	**14.93**

A.4 Pedal Sensing System (PSS)

Surgeon foot pedal interactions convey critical behavioral and operational cues during robot-assisted surgery, including instrument activation patterns, clutching behavior (hand motions decoupled from end-effector movement), and potential indicators of expertise. To capture these signals non-invasively, we developed the *Pedal Sensing System (PSS)*, a compact, low-cost, open source module built around an Arduino-class microcontroller and thin-film force-sensitive resistors (FSRs) (see Figure 14).

Each FSR is mounted in a voltage-divider with a known series resistor R using a custom designed PCB (see Figure 15). Under applied pressure, the FSR resistance R_{FSR} decreases, increasing the divider output V_{out} in Eq. (6), which is sampled by the microcontroller's analog-to-digital converter.

$$V_{\mathrm{out}} = \frac{R}{R + R_{\mathrm{FSR}}} V_{CC}. \tag{6}$$

A brief per-pedal calibration establishes voltage thresholds for robust binary actuation detection at the start of the sensing. Following the initial calibration, we further refine this value through data-driven optimization. Using a subset of labeled ground-truth pedal data, we perform a hyperparameter search over possible voltage thresholds to identify the

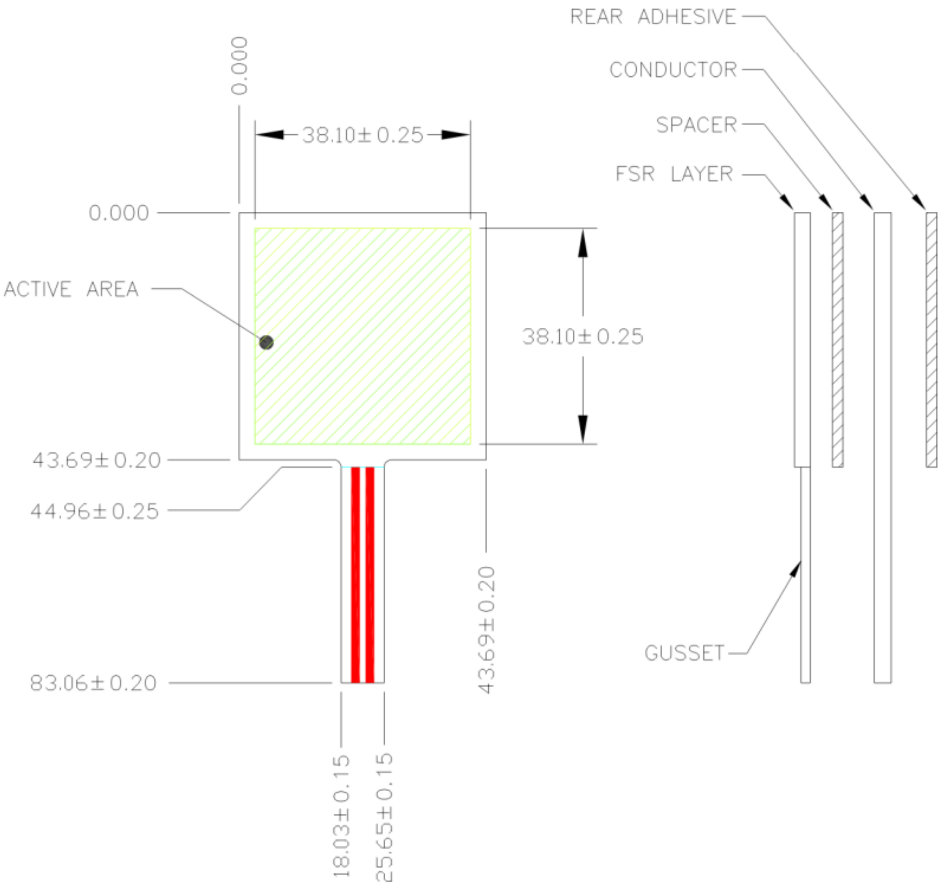


Figure 14: Dimensions of the FSR utilized for the PSS in mm Int [2010].

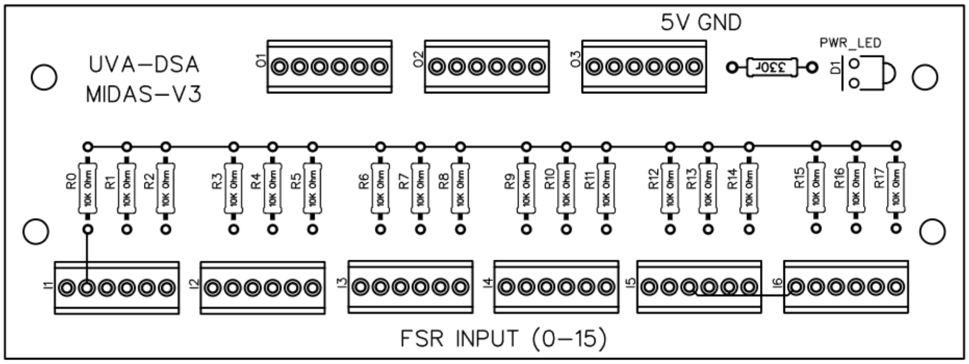


Figure 15: Custom designed PCB schematic for interfacing between microcontroller and FSR sensors.

value that maximizes detection performance. The resulting optimal threshold is then applied to the dataset to evaluate overall PSS accuracy.

Sensors are affixed directly to the surgeon's console pedal surfaces, enabling continuous monitoring without modifying console hardware or perceptibly altering tactile feedback. PSS supports reading up to 9 pedals and records each pedal state along with time-stamped V_{out} values for detailed analysis.

Signals are time-stamped and logged by the MiDAS acquisition system at a specified sampling rate, supporting seamless integration into existing workflows while preserving natural surgeon interaction and providing a high-value stream for behavioral and skill analysis.

B Downstream Task Evaluation: Gesture Recognition

B.1 Peg Transfer Ablation Study - Raven-II

Dataset	Model	Modality	Features	Acc	F1	Prec.	Rec.
Raven-II Peg Transfer (15 trials)	**MTRSAP** Weerasinghe et al. [2024]	PSM	pos., ori., grasp.	0.72	0.75	0.72	0.71
		PSM-NG	pos., ori	0.61	0.58	0.61	0.60
		PSM	**pos., ori., vel., grasp.**	**0.88**	**0.87**	**0.88**	**0.87**
		MTM	pos., ori., grasp.	0.77	0.78	0.77	0.76
		MTM-NG	pos., ori.	0.48	0.47	0.49	0.46
		MTM	pos., ori., vel., grasp.	0.87	0.87	0.90	0.86
		EmHT	pos.	0.68	0.67	0.68	0.67
		EmHT	pos., ori.	0.73	0.75	0.73	0.74
		EmHT	pos., vel.	0.84	0.84	0.85	0.84
		EmHT	**pos., ori., vel.**	**0.87**	**0.86**	**0.87**	**0.86**
		EmHT-NC	pos., ori., vel., no-clutch	0.78	0.77	0.79	0.77
		Image†	RGB (R)	0.39	0.42	0.39	0.39
		Image†	RGB (D)	0.48	0.48	0.48	0.48
		HandKP	pos., vel.	0.40	0.38	0.41	0.39
		HandKP, Image	pos., vel., RGB (R)	0.25	0.16	0.19	0.24
		HandKP, Image	pos., vel., RGB (D)	0.38	0.33	0.36	0.37
	MS-TCN++ Farha and Gall [2019]	PSM	pos., ori., grasp.	0.74	0.76	0.75	0.74
		PSM	pos., ori., vel., grasp.	0.85	0.84	0.86	0.86
		MTM	pos., ori., grasp.	0.76	0.78	0.76	0.74
		MTM	**pos., ori., vel., grasp.**	**0.86**	**0.86**	**0.86**	**0.87**
		EmHT	pos.	0.55	0.53	0.55	0.56
		EmHT	pos., ori	0.65	0.63	0.68	0.68
		EmHT	pos., vel.	0.74	0.74	0.77	0.76
		EmHT	**pos., ori., vel.**	**0.79**	**0.78**	**0.80**	**0.81**
		Image†	RGB (R)	0.44	0.50	0.45	0.43
		Image†	RGB (D)	0.40	0.47	0.41	0.44
		HandKP	pos., vel.	0.33	0.29	0.27	0.37
		HandKP, Image	pos., vel., RGB (R)	0.33	0.32	0.32	0.36
		HandKP, Image	pos., vel., RGB (D)	0.49	0.51	0.49	0.46
DESK Madapana et al. [2019] Peg Transfer (48 trials)	***MTRSAP** Weerasinghe et al. [2024]	PSM	pos., vel., ori., grasp.	0.91	0.91	0.92	0.92
		Image	RGB (R)	0.76	0.77	0.77	0.76
	***MS-TCN++** Farha and Gall [2019]	PSM	pos., vel., ori., grasp.	0.92	0.93	0.92	0.92
		Image	RGB (R)	0.73	0.77	0.75	0.75

Table 13: Raven-II Peg Transfer Gesture Recognition using Different Modalities. Best per model in **bold**.† Image-only modalities underperform due to limited data. * Same baselines evaluated on a larger dataset for comparison. *Abbreviations: pos.=position; vel.=velocity; ori.=orientation; gasp.=grasper angle; NG=No Grasper State Information; NC=No Clutch State Information; (R)=ResNet-50; (D)=DINOv2.*

B.1.1 Feature Ablation

For a comprehensive evaluation of the impact of different features to the performance of gesture recognition, we performed an ablation study on key kinematic features and robot states including position, orientation, and velocity, grasper state and clutch state (see Table 13). We first incorporate clutch state information, which marks periods when surgeon hand motion is decoupled from PSM movement. Excluding these clutching windows during training and inference removes kinematically irrelevant segments and leads to notable performance gains for EmHT, F1 improves from 0.77 to 0.86 demonstrating that clutch state is a valuable auxiliary signal. Given these benefits, clutch state is included for all remaining modalities.

We then evaluated the significance of grasper (jaw) state within the kinematic stream in both PSM and MTM. Removing this signal, an indicator of the surgeon's pinching action caused **macro-F1 to drop significantly from 0.78 to 0.47 in MTM performance and 0.75 to 0.58 in PSM performance**, highlighting its discriminative value. Notably, EmHT (two fingers per manipulator) indirectly captures the pinching motion, providing a proxy for grasper state and partially mitigating the loss when explicit grasper signals are unavailable as discussed previously in EmHT correlation analysis.

We further evaluate feature ablations (position, orientation, and velocity) and find that incorporating velocity consistently yields substantial gains across kinematic modalities, highlighting its importance for gesture recognition.

B.1.2 Visual Feature Adaptation and Pretraining Strategies

In Raven-II Peg Transfer experiments, adding vision features (ResNet-50 Wang and Majewicz Fey [2018], DINOv2 Oquab et al. [2023]) initially failed to surpass kinematic baselines. We attribute this to the low-data regime: high-dimensional encoders trained on few clips overfit appearance rather than gesture discriminative motion; the supervision-to-dimension mismatch (clip-level labels vs. millions of visual DoF) worsens this, and ImageNet to surgery domain shift Deng et al. [2009] (illumination, background and surgical instruments) further limits transfer.

To mitigate the low-data regime, we first pretrained ResNet-50 on the in-domain DESK dataset Madapana et al. [2019], which yielded a 15–30% relative improvement, demonstrating the clear benefit of domain-adaptive pretraining. Replacing ResNet-50 with DINOv2 features provided further gains, consistent with the strength of large-scale self-supervised vision transformers that capture more semantically expressive and globally contextualized representations. Overall, our results highlight that (i) domain-adaptive pretraining is particularly important for CNN backbones, and (ii) foundation-model features such as DINOv2 offer an even stronger initialization under limited supervision, especially when combined with motion cues.

B.2 Suturing Ablation - da Vinci

For the Suturing task, we report per-class gesture recognition performance alongside class distribution statistics to better characterize model performance and generalization across gestures (Figure 16). The best performing model MTRSAPWeerasinghe et al. [2024] has generalized well over the gesture classes except for G5-Reach Suture which has a significantly low number of samples for training.

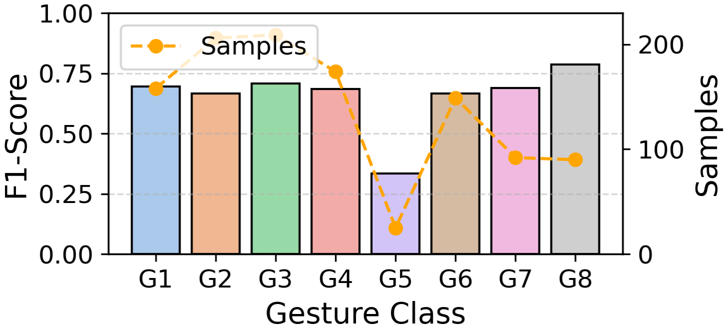
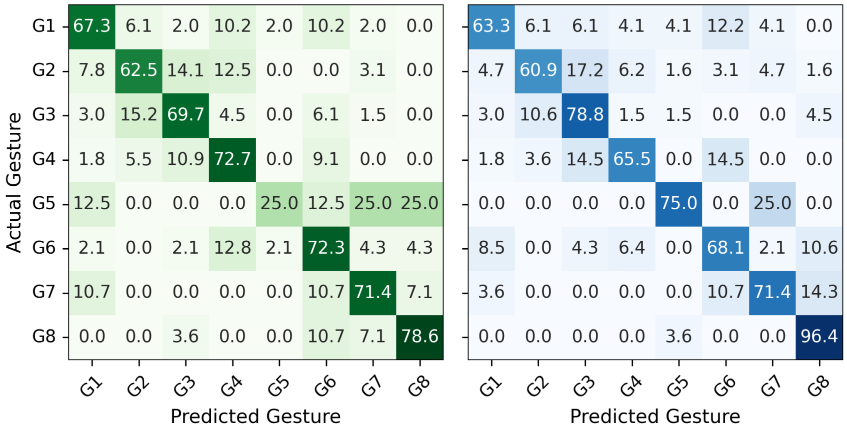


Figure 16: Per-class performance and class distribution for the Suturing dataset.

Moreover, Figure 17 presents confusion matrices for visualization of how the best performing models recognize gestures and which gestures are harder to distinguish. For example, on the da Vinci dataset, fusing EmHT with image features produces clear gains for gestures emphasizing fine needle control (G8-Square Knot and Cinch, G3-Push Needle Through Tissue) and imbalanced classes with fewer samples (G5), but lowers accuracy for visually similar pulling motions (G4-Pull Needle out of Tissue, G6-Pull Suture) as the hand movements are closely aligned kinematically.

Beyond per-class analysis, Table 14 highlights several key modality trends. First, **Image-only** models show moderate performance, with DINOv2 features outperforming ResNet-50 (Acc: 0.69 vs. 0.55 for MTRSAP), reflecting the benefit of foundation-model representations under low-data conditions. Second, **EmHT alone** performs competitively (F1: 0.60 with MTRSAP), indicating that external hand-motion trajectories capture meaningful cues for suturing despite the complexity of real surgical tissue interactions. Third, **fusing EmHT with visual features** yields the strongest results



(a) Image Features

(b) EmHT + Image Features

Figure 17: Confusion matrices for modality fusion on the da Vinci Xi using MTRSAP Weerasinghe et al. [2024].

(Acc: 0.71, F1: 0.70 for MTRSAP), demonstrating that motion dynamics and visual context are complementary for recognizing fine-grained suturing gestures.

In contrast, **HandKP** features derived from RGB-D hand keypoint tracking perform notably worse in isolation (F1: 0.35) and offer limited gains when fused with image features. This is expected, as keypoints are frequently missed when the hands leave the camera's field of view, leading to fragmentary motion information.

Overall, the Suturing ablations show that (i) **DINOv2 image features** provide the strongest visual baseline, (ii) **EmHT** is an effective non-invasive proxy for robot-side motion, and (iii) **multimodal fusion** yields the most robust performance across gesture types, especially under class imbalance and variable tissue interactions.

Dataset	Model	Modality	Features	Acc	F1	Prec.	Rec.
da Vinci Xi Suturing (17 trials)	**MTRSAP**Weerasinghe et al. [2024]	Image	RGB (R)	0.55	0.57	0.57	0.56
		Image	RGB (D)	0.69	0.67	0.65	0.65
		EmHT	pos., ori., vel.	0.64	0.60	0.63	0.61
		EmHT, Image	pos., ori., vel., RGB (R)	0.60	0.57	0.63	0.57
		EmHT, Image	**pos., ori., vel., RGB (D)**	**0.71**	**0.70**	**0.70**	**0.69**
		HandKP	pos., vel.	0.40	0.35	0.37	0.36
		HandKP, Image	pos., vel., RGB (R)	0.50	0.46	0.49	0.47
		HandKP, Image	pos., vel., RGB (D)	0.58	0.58	0.58	0.57
	MS-TCN++Farha and Gall [2019]	Image	RGB (R)	0.53	0.44	0.44	0.43
		Image	RGB (D)	0.57	0.57	0.53	0.51
		EmHT	pos., ori., vel.	0.57	0.50	0.50	0.53
		EmHT, Image	pos., ori., vel., RGB (R)	0.58	0.51	0.51	0.52
		EmHT, Image	**pos., ori., vel., RGB (D)**	**0.63**	**0.56**	**0.64**	**0.63**
		HandKP	pos., vel.	0.43	0.36	0.39	0.37
		HandKP, Image	pos., vel., RGB (R)	0.48	0.42	0.46	0.45
		HandKP, Image	pos., vel., RGB (D)	0.58	0.54	0.58	0.54
SAR-RARP50 Psychogyios et al. [2023] Suturing (50 trials)	**ASFormer Yi et al. [2021]**	Image	RGB	0.82†	–	–	–
	MViT Fan et al. [2021]	Image	RGB	0.79†	–	–	–
SAIS Kiyasseh et al. [2023] Suturing (78 trials)	**Vision Transformer** Kiyasseh et al. [2023]	Image	RGB	–	0.50†	–	–

Table 14: Modality ablation for Suturing on the da Vinci Xi, with features per modality. Best results per model in **bold**. *Abbreviations: pos.=position; vel.=velocity; ori.=orientation; (R)=ResNet-50; (D)=DINOv2; †=As reported or estimated from published results.*

References

Ziheng Wang and Ann Majewicz Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 13(12):1959–1970, 2018. ISSN 1861-6429.

Andrew J Hung, Jian Chen, and Inderbir S Gill. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA surgery*, 153(8):770–771, 2018.

Arnaud Huaulmé, Kanako Harada, Quang-Minh Nguyen, Bogyu Park, Seungbum Hong, Min-Kook Choi, Michael Peven, Yunshuang Li, Yonghao Long, Qi Dou, Satyadwyoom Kumar, Seenivasan Lalithkumar, Ren Hongliang, Hiroki Matsuzaki, Yuto Ishikawa, Yuriko Harai, Satoshi Kondo, Manoru Mitsuishi, and Pierre Jannin. PEg TRAnsfer Workflow recognition challenge report: Do multimodal data improve recognition? *Computer Methods and Programs in Biomedicine*, 236:107561, 2023.

Mohammad Samin Yasar and Homa Alemzadeh. Real-time context-aware detection of unsafe events in robot-assisted surgery. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 385–397. IEEE, 2020.

Jenna C Allers, Ahmed A Hussein, Nabeeha Ahmad, Lora Cavuoto, Joseph F Wing, Robin M Hayes, Nobuyuki Hinata, Ann M Bisantz, and Khurshid A Guru. Evaluation and impact of workflow interruptions during robot-assisted surgery. *Urology*, 92:33–37, 2016.

Francesco Cepolina and Roberto P Razzoli. An introductory review of robotically assisted surgical systems. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 18(4):e2409, 2022.

Shunsuke Miyamoto, Tomoya Hatayama, Hiroyuki Shikuma, Kazuma Yukihiro, Kyohsuke Iwane, Ryo Tasaka, Yuki Kohada, Takafumi Fukushima, Kenshiro Takemoto, Miki Naito, Kohei Kobatake, Yohei Sekino, Hiroyuki Kitano, Kenichiro Ikeda, Keisuke Goto, Akihiro Goriki, Keisuke Hieda, and Nobuyuki Hinata. Robotic urologic applications of the hinotori™ surgical robot system. *Asian Journal of Urology*, 12(2):162–168, 2025. ISSN 2214-3882. Robotics, artificial intelligence, telepresence, and telesurgery: The future of urology.

Blake Hannaford, Jacob Rosen, Diana W Friedman, Hawkeye King, Phillip Roan, Lei Cheng, Daniel Glozman, Ji Ma, Sina Nia Kosari, and Lee White. Raven-ii: an open platform for surgical robotics research. *IEEE Transactions on Biomedical Engineering*, 60(4):954–959, 2012.

Naveen Madapana, Md Masudur Rahman, Natalia Sanchez-Tamayo, Mythra V Balakuntala, Glebys Gonzalez, Jyothsna Padmakumar Bindu, LN Vishnunandan Venkatesh, Xingguang Zhang, Juan Barragan Noguera, Thomas Low, et al. Desk: A robotic activity dataset for dexterous surgical skills transfer to medical robots. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6928–6934. IEEE, 2019.

Xiao Li, Homa Alemzadeh, Daniel Chen, Zbigniew Kalbarczyk, Ravishankar K Iyer, and Thenkurussi Kesavadas. A hardware-in-the-loop simulator for safety training in robotic surgery. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5291–5296. IEEE, 2016.

Yun-Hsuan Su, Adnan Munawar, Anton Deguet, Andrew Lewis, Kyle Lindgren, Yangming Li, Russell H. Taylor, Gregory S. Fischer, Blake Hannaford, and Peter Kazanzides. Collaborative robotics toolkit (CRTK): Open software framework for surgical robotics research. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 48–55, 2020.

Peter Kazanzides, Zihan Chen, Anton Deguet, Gregory S Fischer, Russell H Taylor, and Simon P DiMaio. An open-source research kit for the da vinci® surgical system. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 6434–6439. IEEE, 2014.

Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.

Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, et al. Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 212:106452, 2021.

Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamín Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3, 2014.

Pascal Hansen, Ji Woong Brian Kim, Antony Goldenberg, Juo Tung Chen, Yuanzhe Amos Li, Anton Deguet, Brandon White, De Ru Tsai, Richard Cha, Jeffrey Jopling, et al. Imitatecholec: A multimodal dataset for long-horizon imitation learning in robotic cholecystectomy. *Scientific Data*, 2026.

Mahtab J Fard, Sattar Ameri, R Darin Ellis, Ratna B Chinnam, Abhilash K Pandya, and Michael D Klein. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1):e1850, 2018.

Ziheng Wang and Ann Majewicz Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International journal of computer assisted radiology and surgery*, 13(12):1959–1970, 2018.

Kay Hutchinson, Zongyu Li, Leigh A Cantrell, Noah S Schenkman, and Homa Alemzadeh. Analysis of executional and procedural errors in dry-lab robotic surgery experiments. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 18(3):e2375, 2022.

Adam Goldbraikh, Tomer Volk, Carla M Pugh, and Shlomi Laufer. Using open surgery simulation kinematic data for tool and gesture recognition. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–15, 2022.

Keshara Weerasinghe, Seyed Hamid Reza Roodabeh, Kay Hutchinson, and Homa Alemzadeh. Multimodal transformers for real-time surgical activity prediction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13323–13330. IEEE, 2024.

Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.

Ki-Hwan Oh, Leonardo Borgioli, Alberto Mangano, Valentina Valle, Marco Di Pangrazio, Francesco Toti, Gioia Pozza, Luciano Ambrosini, Alvaro Ducas, Miloš Žefran, Liaohai Chen, and Pier Cristoforo Giulianotti. Comprehensive robotic cholecystectomy dataset (crcd): Integrating kinematics, pedal signals, and endoscopic videos. In *2024 International Symposium on Medical Robotics (ISMR)*, pages 1–7, 2024.

Nasseh Hashemi, Morten Bo Søndergaard Svendsen, Flemming Bjerrum, Sten Rasmussen, Martin G Tolsgaard, and Mikkel Lønborg Friis. Acquisition and usage of robotic surgical data for machine learning analysis. *Surgical Endoscopy*, 37(8):6588–6601, 2023.

Kay Hutchinson. *Fine-Grained Activity Modeling, Recognition, and Error Analysis in Robot-Assisted Surgery*. PhD thesis, University of Virginia, Charlottesville, VA, USA, 2023.

Jože Guna, Grega Jakus, Matevž Pogačnik, Sašo Tomažič, and Jaka Sodnik. An analysis of the precision and reliability of the Leap Motion sensor and its suitability for static and dynamic tracking. *Sensors*, 14(2):3702–3720, 2014.

Yonjae Kim, Peter CW Kim, Rebecca Selle, Azad Shademan, and Axel Krieger. Experimental evaluation of contact-less hand tracking systems for tele-operation of surgical tasks. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3502–3509. IEEE, 2014.

Adam Goldbraikh, Omer Shubi, Or Rubin, Carla M Pugh, and Shlomi Laufer. Kinematic data-based action segmentation for surgical applications. *arXiv preprint arXiv:2303.07814*, 2023.

Northern Digital Inc. 3d guidance: Quality oem tracking solution. `https://www.ndigital.com/electromagnetic-tracking-technology/3d-guidance/`, 2024. Accessed: 2025-09-30.

Stereolabs Inc. Zed mini stereo camera – product page. `https://www.stereolabs.com/store/products/zed-mini`. Accessed: 2025-11-04.

OBS Project. Obs studio. `https://obsproject.com/`, 2025. Open-source software.

Adafruit Industries. Using an fsr | force sensitive resistor (fsr), July 2012. URL `https://learn.adafruit.com/force-sensitive-resistor-fsr/using-an-fsr`. Accessed: 2025-10-31.

E Matt Ritter and Daniel J Scott. Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery. *Surgical innovation*, 14(2):107–112, 2007.

Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6/19/10.

Richard H. Feins. *Real Tissue Robotic Simulation: The KindHeart Simulators*, pages 105–109. Springer International Publishing, Cham, 2021. ISBN 978-3-030-53594-0.

Kay Hutchinson, Ian Reyes, Zongyu Li, and Homa Alemzadeh. Compass: a formal framework and aggregate dataset for generalized surgical procedure modeling. *International Journal of Computer Assisted Radiology and Surgery*, 18 (12):2143–2154, 2023.

Dani Kiyasseh, Runzhuo Ma, Taseen F Haque, Brian J Miles, Christian Wagner, Daniel Donoho, Anima Anandkumar, and Andrew J Hung. A vision transformer for decoding surgeon activity from surgical videos. *Nature Biomedical Engineering*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Luca Zappella, Benjamin Bejar, Gregory Hager, and Rene Vidal. Surgical gesture classification from video and kinematic data. *Medical image analysis*, 17(7):732–745, 2013.

Chang Shi, Yi Zheng, and Ann Majewicz Fey. Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8017–8024. IEEE, 2022.

Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

Marcel Pfister, Jaw-Chyng L Lue, Francisco R Stefanini, Paulo Falabella, Laurie Dustin, Michael J Koss, and Mark S Humayun. Comparison of reaction response time between hand and foot controlled devices in simulated microsurgical testing. *BioMed research international*, 2014(1):769296, 2014.

Chiho Lim, Juan Antonio Barragan, Jason Michael Farrow, Juan P Wachs, Chandru P Sundaram, and Denny Yu. Physiological metrics of surgical difficulty and multi-task requirement during robotic surgery skills. *Sensors*, 23(9): 4354, 2023.

Jing Yang, Juan Antonio Barragan, Jason Michael Farrow, Chandru P. Sundaram, Juan P. Wachs, and Denny Yu. An adaptive human-robotic interaction architecture for augmenting surgery performance using real-time workload sensing—demonstration of a semi-autonomous suction tool. *Human Factors*, 66(4):1081–1102, 2024. PMID: 36367971.

Radu Horaud and Fadi Dornaika. Hand-eye calibration. *The International Journal of Robotics Research*, 14(3):195–210, 1995.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172, 2019. Available at `https://arxiv.org/abs/1906.08172`.

FSR 406 Data Sheet. Interlink Electronics, 10 2010. Rev. A.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.