2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SurgSync: Time-Synchronized Multi-Modal Data Collection Framework and Dataset for Surgical Robotics

*Haoying Zhou[1,2], *Chang Liu[2], Yimeng Wu[2], Junlin Wu[2,3], Zijian Wu[4], Yu Chung Lee[4], Sara Martuscelli[5], Septimiu E. Salcudean[4], Gregory S. Fischer[1] and Peter Kazanzides[2,3]

***Abstract*—Most existing robotic surgery systems adopt a human-in-the-loop paradigm, often with the surgeon directly teleoperating the robotic system. Adding intelligence to these robots would enable higher-level control, such as supervised autonomy or even full autonomy. However, artificial intelligence (AI) requires large amounts of training data, which is currently lacking. This work proposes SurgSync, a multi-modal data collection framework with offline and online synchronization to support training and real-time inference, respectively. The framework is implemented on a da Vinci Research Kit (dVRK) and introduces (1) dual-mode (online/offline-matching) synchronized recorders, (2) a modern stereo endoscope to achieve image quality on par with clinical systems, and (3) additional sensors such as a side-view camera and a novel capacitive contact sensor to provide ground truth contact data. The framework also incorporates a post-processing toolbox for tasks such as depth estimation, optical flow, and a practical kinematic reprojection method using Gaussian heatmap. User studies with participants of varying skill levels are performed with ex-vivo tissue to provide clinically realistic data, and a network for surgical skill assessment is employed to demonstrate utilization of the collected data. Through the user study experiments, we obtained a dataset of 214 validated instances across multiple canonical training tasks. All software and data are available at surgsync.github.io.**

I. INTRODUCTION

Robot-assisted surgery (RAS) has revolutionized the field of medical science by providing surgeons with enhanced dexterity, advanced visualization, and precision when performing clinical procedures over the past two decades. The success of the da Vinci® Surgical System (dVSS, Intuitive Surgical Inc. Sunnyvale, CA) is the epitome of this revolution [1].

In the research domain, high-quality, well-annotated datasets are foundational to progress in artificial intelligence (AI) applications [2] for RAS [3]. They enable data-driven perception, modeling and control, spanning instrument tracking [4]–[9], tissue interaction understanding [10]–[12], skill assessment [13]–[16], and surgery automation [17]–[21].

This work was supported in part by NSF AccelNet awards OISE-1927275 and OISE-1927354.

*These authors contributed equally to this work.
[1]Department of Robotics Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. Emails: hzhou6, gfischer@wpi.edu
[2]Laboratory for Computational Sensing and Robotics, Johns Hopkins University, Baltimore, MD, USA.
[3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. Email: pkaz@jhu.edu
[4] Robotics and Control Laboratory, the University of British Columbia, Vancouver, Canada.
[5] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

Despite this momentum, robotics applications, especially for surgical scenarios, face a shortage of high-quality training and validation data due to highly diverse data distribution and the expensive cost of data collection in the physical world. Even though generating synthetic data for training [22]–[28] can mitigate this problem through sim-to-real transfer, the sim-to-real gap limits the complexity of procedures and overall performance. Therefore, there remains a need for real-world datasets to ensure effective evaluation.

Many available datasets in the surgical robotics domain suffer from three practical limitations: (i) weak or inconsistent time alignment across sensing modalities, which obscures cause-and-effect and degrades sequence models; (ii) legacy imaging pipelines that limit visual fidelity and downstream vision performance; and (iii) narrow task coverage and post-collection tooling that constrain reproducibility and reuse. These gaps are especially consequential for systems like the da Vinci Research Kit (dVRK, also known as dVRK Classic) [29] and dVRK-Si [30], where fine motor actions, bi-manual coordination, and tissue dynamics evolve on sub-second timescales and must be captured coherently across vision and robot states.

To address these challenges, we present an open-source data collection framework for surgical robotic systems, such as dVRK Classic and dVRK-Si, with the following contributions:

- **Time-synchronization design pattern:** two synchronized recorders (online/offline-matching) for smooth and continuous teleoperation recording, acknowledging time synchronization as a first-class design constraint;
- **Upgraded imaging stack:** integration of a modern chip-on-tip endoscope (Cornerstone Robotics (CSR) Ltd., Hong Kong, China) with dVRK-Si to enable high-performance imaging stack;
- **Tool-tissue contact ground-truth sensing:** a capacitive contact sensor, interfaced via a digital input of the dVRK controller, for seamless acquisition of tool-tissue contact ground truth on ex-vivo tissues;
- **Post-collection processing toolbox:** a configurable and extensible post-collection toolbox for better reusability;
- **User-study dataset:** user studies for data collection, including multiple practical training procedures performed on phantoms and ex-vivo tissues, such as peg transfer, tissue manipulation, suturing and dissection.

In addition, we validate the usability of our dataset via implementing a state-of-the-art skill assessment algorithm

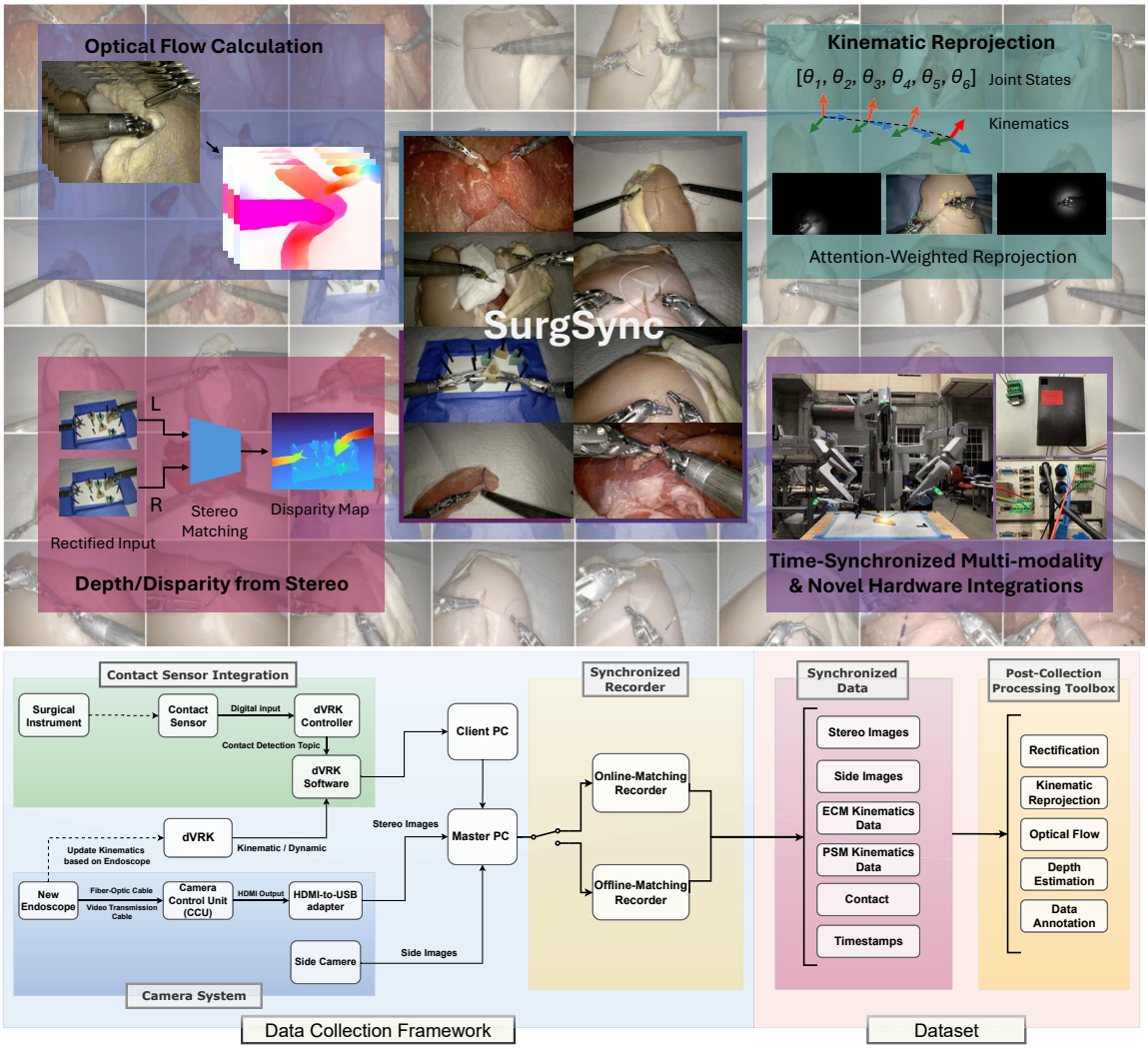

Fig. 1. Overview of the proposed SurgSync dataset and data collection framework. We collect 214 recordings while performing multiple canonical training tasks on ex-vivo tissues (primarily) and phantoms. We also implement post-collection processing using our toolbox. Multiple input modalities, including visual, kinematics/dynamics, tool-tissue contact, event/phase description, are provided for further algorithm training.

[15] on the suturing-task subset of our data.

Our proposed framework primarily communicates via Robot Operating System (ROS) and can be extended to general surgical robots supported within the ROS ecosystem. Furthermore, the cross-platform validation demonstrates that the framework can be used by multiple groups to create a combined, diverse, large-scale dataset. An overview of the system and dataset are shown in Fig. 1.

II. Related Work

The surgical robotics domain has been facing a lack of high-quality, well-annotated large-scale public datasets. Researchers have devoted substantial efforts to generate synthetic data through simulation [22]–[28], [31]–[33], however, a persistent trade-off remains among photorealism, physically faithful tool-tissue interactions and task complexity. Therefore, real-world data remain essential for training and rigorous evaluation.

In 2014, Gao et al. introduced JIGSAWS [34], [35], the first open-source dataset for surgical gesture recognition. It catalyzed AI research in RAS [36] yet exhibits known limitations [37], including suboptimal image quality and visual fidelity due to the legacy image pipeline. Also in 2014, we introduced the da Vinci Research Kit (dVRK) [29], an open-source research platform derived from the dVSS, which enabled collection of additional datasets on phantoms for peg transfer [21], tool retraction/palpation [11], pattern cutting [38] and other related training tasks [39]. However, both JIGSAWS and these dVRK datasets were collected with phantoms, which may not capture the dynamics of ex-vivo or in-vivo tool-tissue interactions.

Beyond these efforts, the EndoVis challenges have released public datasets over several years [6]–[8], providing benchmarks for endoscopic perception and segmentation tasks. Various multi-modal endoscopic datasets [40] have also been released for Large Language Model (LLM) investigations. All aforementioned datasets share one or more constraints: (i) suboptimal time alignment across different modalities, (ii) narrow task coverage, (iii) absence of instrument/robot kinematic data, (iv) missing camera parameters for further calibration, and (v) lack of multi-view image information.

Finally, our previous work [9] investigated advanced annotation using fluorescence-based approaches, but logged only sparse, discrete events, which limits temporal resolution and can obscure critical dynamics information.

III. Methodology

A. Synchronized Recorder

We design and implement two practical synchronized recorders in modern C++, which aim to record temporally aligned data streams across multiple sensing modalities. Specifically, they synchronize stereo (or mono) video streams with kinematic data from the patient-side manipulators (PSMs) and endoscopic camera manipulator (ECM) of the dVRK, including both measured and desired (also known as setpoint) states. In addition, the contact sensor signals are seamlessly integrated within the dVRK data stream and recorded with the same time base. We offer two operation modes for the synchronized recorders so that the user can choose between online or offline time-matching approaches. Both recorders operate on smooth and continuous teleoperation, preserving fine-grained temporal context for sequence models, dynamics and interaction analysis. All communications are based on ROS.

1) Online-Matching Recorder: The design enforces strict time synchronization using multi-threading: only samples that fall within a user-defined time tolerance are admitted. This yields slightly uneven inter-sample intervals (irregular Δt), but retains the natural continuity of smooth teleoperation segments and avoids label/feature drift. In our study, a time tolerance of 10 ms is selected to ensure both tight time alignment and consecutive recorder output. The choice of the time tolerance depends on task requirements and hardware I/O performance. This design can be used for real-time scenarios.

2) Offline-Matching Recorder: Our offline-matching approach decouples recording from time alignment to maximize the recording system efficiency. Therefore, this recorder produces synchronized datasets in two stages: (i) a lightweight recorder logs camera streams to videos and raw kinematic streams to binary files with minimal processing; (ii) an offline post-processing pipeline reconstructs a fixed-rate frame sequence and, for each frame, gathers the k closest samples using nearest-timestamp lookup for subsequent interpolation (we select $k = 1$ for simplicity). Compared to the online-matching recorder (which pairs visual and kinematic data in real time), this two-stage design avoids tolerance-based dropping of data during capture, yielding a higher throughput and uniform intervals between synchronized data packets at the cost of requiring more storage and substantial time for post-collection time-matching and interpolation.

This offline-matching recorder is intended for data collection where real-time data acquisition is non-essential, for example, building large training datasets for robot policy learning. By recording every camera frame at the target FPS and interpolating kinematics offline using a configurable rule, the pipeline eliminates tolerance-based drops at capture

Algorithm 1 Online-Matching Recorder

procedure OnlineMatchRecorder
- **Initialize Subscribers**
  - Subscribe to video streams based on user preferences
  - **For each** enabled arm (PSMs or ECM):
    - Subscribe to `<arm>/measured_js/cp/cv`
    - Subscribe to `<arm>/setpoint_js/cp`
    - **If** arm is PSM: subscribe to jaw streams, contact sensors
- **Data Structures**
  - Buffers: queues for images, kinematics (per arm)
  - `SyncedQueue`: queue of synchronized data packets
  - Global state: latest setpoint & jaw values (per arm)
- **Sync Thread**
  - **Loop** until cut-off requested from keyboard input:
    - **If** `SyncedQueue` full $\rightarrow$ drop oldest images, continue
    - **If** image buffers empty $\rightarrow$ wait
    - Retrieve image reference timestamp
    - **For each** arm:
      - `GetClosestFromQueue`(kin_buffer, ref_stamp)
    - Construct `SyncedPacket` including reference time--stamp, images and kinematic information
    - Enqueue packet to `SyncedQueue`
    - Pop used image(s) from buffer
    - Notify writer threads
- **Writer Threads** (default pool size: $N = 4$)
  - **Loop:**
    - Wait on `SyncedQueue`, pop one packet
    - Create temporary folder `<timestamp>`
    - Save the images as PNGs
    - Save the arm kinematics to JSON files
    - (Optional) Save data to HDF5 and convert later
- **Shutdown and Cleanup (Press 'q' on the keyboard)**
  - Obtain the cut-off timestamp t_{end}
  - Release buffers until the reference timestamp passes t_{end}
  - Remove incomplete folders
  - `ReformatDataStorage()` into final dataset layout

end procedure

Algorithm 2 Offline-Matching Recorder

procedure DetachedRecorder
- **Initialize Directory**
  - Create `<run_id>` with subdirs: `kin/`, `meta/`
- **Initialize Subscribers**
  - Subscribe to video streams based on user preferences
  - **For each** enabled arm (PSMs or ECM):
    - Subscribe to `<arm>/measured_js/cp/cv`
    - Subscribe to `<arm>/setpoint_js/cp`
    - **If** PSM: subscribe to jaw streams, contact sensors
- **Video Writer** (fixed-rate)
  - Create `VideoStreamRecorder(s)` for left/right[/side]
  - Record frames at pre-defined FPS
  - **If** ahead of schedule:
    - Wait until next frame's expected timestamp
  - Image callbacks push frames to corresponding recorder(s)
- **Kinematic Writer** (binary)
  - **For each** arm/topic callback:
    - Append to `FastBinWriter` file in `kin/`
- **Runtime Control**
  - Start AsyncSpinner for callbacks
  - Launch key listener; press 'q' $\rightarrow$ stop recording
- **Shutdown**
  - Stop recorders; flush and close all writers
  - Write `start_times.json` and `end_times.json`

end procedure

procedure Detached Post-Collection Matching
- Decouple the videos into independent image frames
- Convert binary kinematic files to readable JSON files

end procedure

and produces a uniformly sampled training set. It is the more suitable choice for offline learning workflows, where deterministic post-processing is preferable to on-the-fly data matching.

B. Contact Sensor Integration

Our framework supports the integration of custom sensors. For example, we incorporated a contact sensor implemented with an Arduino UNO Rev3 and the Capacitive Sensing Library [41]. The assembled hardware prototype and schematics are shown in Fig. 2.

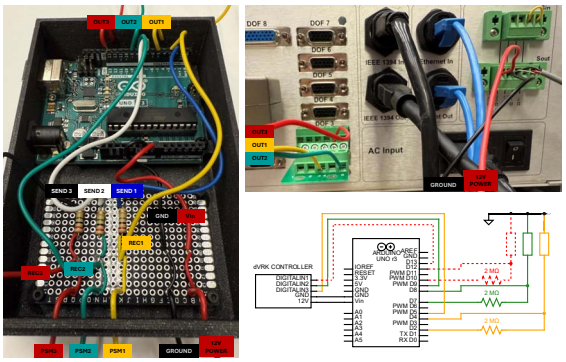

Fig. 2. Hardware setup and schematics. The Arduino and the proto-board are shown on the left, the schematics of the electronics on the bottom right, and the connection to the dVRK controller on the top right.

For monopolar and bipolar instruments, the connection to the sensor is achieved by wrapping a wire around their connectors; for non-electrosurgical (non-polar) instruments, we opened the tool housing, and attached a wire to the rotation coupling of one joint with proper strain relief and wire insulation (Fig. 3).

Fig. 3. Instrument wire connection. Non-polar instrument (left), monopolar (center), and bipolar (right). For non-polar instrument, the tool housing is opened and the wire insulation layer is removed for better visualization.

The library computes the capacitance at the receiving pin in arbitrary units; the returned values depend on the chosen resistor and on the material and size of the contacting object. The sensor exhibits the best performance with human tissue, while ex-vivo animal tissue such as chicken breast, thin-sliced beef or pork also provide similarly reliable results. A contact threshold is defined to binarize the signal into contact and non-contact states, which are then transmitted through the Arduino output pin to the digital input of the dVRK controller (Fig. 2, top right). Each digital input is registered in the dVRK software framework by specifying bit ID, FPGA board, and related parameters in a configuration file.

In this study, we used 2 MΩ resistors and a threshold of 205 for the best performance on the ex-vivo tissues used in our experiments. To evaluate the contact sensor performance, we randomly select one instance each from tissue manipulation, suturing and dissection, as shown in section IV-B, and assess detection accuracy. The resulting accuracies are 99.1% for tissue manipulation, 74.3% for dissection and 45.2% for suturing. The misclassifications are primarily attributed to sensor noise, humidity-induced changes in capacitance and intermittent short-circuit events when instruments contacted the same conductive object (e.g., a suturing needle or a small tissue fragment). Those misclassifications are eventually addressed by manual re-annotation using the GUI in section III-D2.

C. Modern Endoscope Integration

We integrate a contemporary chip-on-tip endoscope with the dVRK-Si system to improve visual fidelity and enable frame-accurate alignment with robot kinematics and contact signals. The full imaging pipeline is shown in Fig. 1 (bottom), comprising hardware mounting, signal capture, and ROS integration. Our integration yields higher-quality images compared to the default dVRK-Si scope and enables coherent multi-modal dataset construction.

1) Hardware Integration: The endoscope connects to a clinical-grade Camera Control Unit (CCU) via video and fiber-optic illumination cables. The CCU outputs two 1080p HDMI video streams, which are routed to the host PC using HDMI-to-USB frame capture devices. A compact 3D-printed holder mounts the endoscope onto the dVRK-Si Endoscopic Camera Manipulator (ECM) with coaxial alignment, quick attachment, and strain relief. CAD models and overall system setup are shown in Fig. 4.

2) Software Integration: On the host PC, frames are acquired through the `v4l2src` backend using `gscam` & `gstreamer` and published to ROS image topics. Each frame is timestamped on arrival using the host monotonic clock. This shared time base enables precise temporal alignment during post-processing. The same pipeline supports stereo

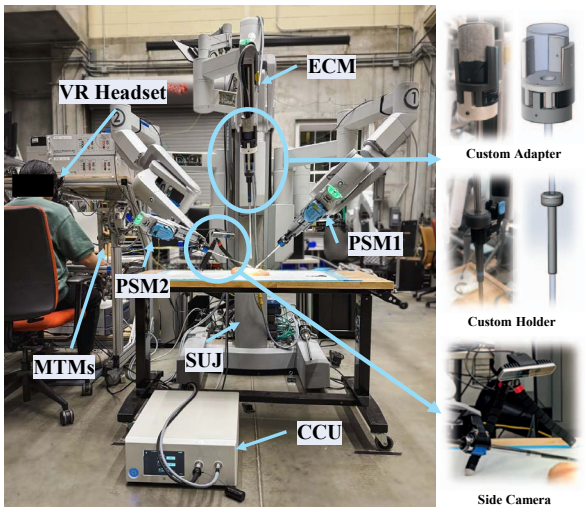

Fig. 4. Overall experimental setup. The endoscope is mounted on the dVRK-Si ECM using a custom holder and adapter.

or side-view cameras through multiple synchronized ROS nodes.

3) Image Quality Comparison: Compared to the legacy dVRK-Si endoscope, our integrated imaging system yields significantly sharper visual frames. Quantitatively, the average Laplacian variance [42] is over 30× higher in our system as shown in Table I, indicating substantially improved details and edge clarity. These improvements benefit downstream perception tasks such as segmentation, optical flow calculation, and depth estimation. Representative frames from each system are shown in Fig. 5.

TABLE I
LAPLACIAN VARIANCE ON SUTURING SEQUENCES.

System	Mean	STD
CSR Endoscope (Ours)	529.48	23.77
dVRK-Si Endoscope	16.93	2.47

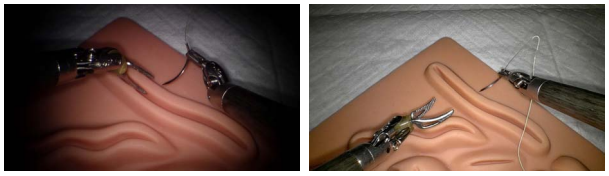

Fig. 5. Representative frames from the dVRK-Si endoscope (left) and our modern CSR endoscope (right). *Brightness settings:* dVRK-Si CCU at 100%, ours at 30%. Both frames are captured with similar camera-phantom distances.

D. Post-Collection Processing Toolbox

Beyond data acquisition, we provide an extensible and highly configurable post-collection processing toolbox that standardizes calibration and data processing. It supports stereo rectification, image resizing, projection of robot kinematics into the endoscopic image frames, and generation of derived modalities, including disparity/depth and optical flow. In addition, it also produces comprehensive annotations for contact detection and event/phase labels.

1) Kinematic Reprojection: We propose a practical approach using a Gaussian heatmap to project tool tip 3D position kinematic information [43] to 2D gray-scale endoscopic images. This relies on a hand-eye calibration [44] to correct for the well-known inaccuracy of the dVRK kinematics [45].

Given the hand-eye calibration and stereo camera parameters, the 3D point $p = (x_p, y_p, z_p)^T$ on the dVRK PSM tool-yaw link (shown in Fig. 6) is projected to the image plane (u_p, v_p) using a pinhole camera model. We can then generate a Gaussian heatmap G centered at (u_p, v_p):

$$G(p_x, p_y) = e^{-\left(\frac{(p_x-u_p)^2}{\sigma_x^2}+\frac{(p_y-v_p)^2}{\sigma_y^2}\right)} \quad (1)$$

where (p_x, p_y) indexes image pixels and σ_x, σ_y control the spread of the heatmap along the image axes.

Eventually, we compute an element-wise product between the rectified grayscale stereo images (I_{gray}) and the heatmap mask G to obtain attention-weighted images ($I_a = G \odot I_{gray}$) that emphasize the region of interest for further interaction analysis. The overall pipeline is shown in Fig. 6.

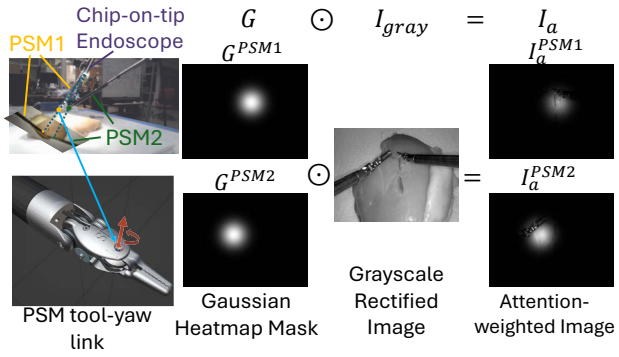

Fig. 6. Kinematic Reprojection Pipeline. This approach projects the Cartesian positions of the PSM tool-yaw link to attention-weighted images.

2) Data Annotation: We developed a custom data annotator with Graphical User Interface (GUI) using PyQt as shown in Fig. 7, enabling users to label contact detection and event/phase description during playback of consecutive frames.

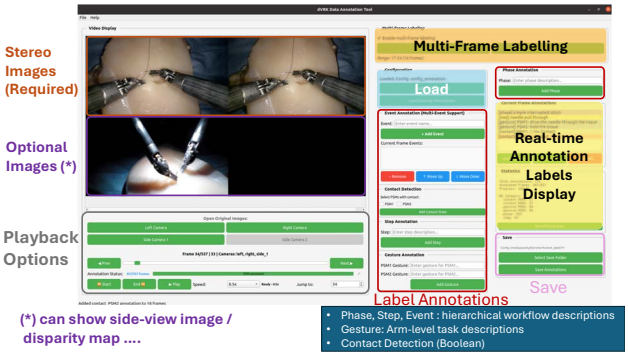

Fig. 7. Data Annotation GUI

3) Depth Estimation: As shown in Fig. 1, we perform disparity estimation using the FoundationStereo model [46] and obtain disparity images. Furthermore, the disparity can be converted to depth via $depth = \frac{f \times b}{disparity}$, where f and b represent the focal length and baseline distance of the stereo endoscope that can be directly obtained from the stereo camera parameters.

4) Optical Flow: As shown in Fig. 1, we also employ the RAFT model [47] to compute dense optical flow between consecutive frames. We then apply a custom magnitude-aware filter to suppress low-magnitude noise for better image output.

IV. EXPERIMENT SETUP

A. System Configuration

Fig. 4 shows the primary system configuration at Johns Hopkins University (JHU). To assess cross-platform generality, we deployed a second configuration on the dVRK Classic at the University of British Columbia (UBC). Both systems share the same camera architecture (stereo endoscope + side-view camera), although the UBC system uses the legacy dVRK Classic endoscope rather than the modern unit and omits the contact sensor. In the UBC setup, cameras stream to

a client workstation, while the I/O-intensive dVRK software framework runs on the master workstation. The hand-eye calibration was not performed at UBC. Both setups use an Intel® RealSense^{TM} RGBD camera as the side-view camera.

Both recorder systems run on workstation-class CPUs: Intel® Xeon(R) W-2245 (JHU) and Intel® Core^{TM} i9-11900 (UBC). All cameras produce raw image streams with a resolution of 1080p and 30 Hz (side-view) or 60 Hz (stereo endoscope) refresh rate. The dVRK framework runs at 1 kHz. At JHU, the offline-matching recorder sustains up to 10 Hz and omitting the side-view camera can increase the rate to 15 Hz.

B. User Study

We conduct a user study to collect data for multiple canonical training tasks on phantoms or ex-vivo tissues, which include but are not limited to:

- Peg transfer (phantoms and gauze in chicken breast)
- Single interrupted suturing practice (chicken breast)
- Tissue manipulation (chicken hearts/breast, beef and pork)
- Dissection following a trace (beef and pork)

13 human subjects (3 female, 10 male) participated in our user study. 4 subjects have limited knowledge of dVRK operation and are considered novice users (N); 5 are familiar with dVRK operation and are considered experienced users (E); 4 are surgeons and are considered professional users (P).

V. Results

A. Cross-Platform Validation

The open-source data collection framework successfully executed on two different setups at two different institutions (JHU and UBC). The UBC setup achieved similar performance using a larger online-matching time tolerance of 100 ms due to the heavy I/O load from the dVRK software framework.

B. Dataset Distribution

We collected 214 instances over multiple training tasks in our user study. The total frame count per instance varies, depending on the task and the user's skill level. Of the 214 validated instances, 9 were collected on the UBC setup and only used the online-matching recorder without the hand-eye calibration. 102 instances were collected using the offline-matching recorder. 96 offline-matching instances were collected within the Intuitive abdominal dome. A brief dataset distribution with respect to training tasks and human subjects' skill level is shown in Table II.

C. Synchronized Recorder Evaluation

To evaluate time-synchronization performance, we randomly selected 4 instances for each recorder from the JHU dataset. The online-matching recordings contain 1523 frames in total and the offline-matching recordings contain 4876 frames. For each frame, excluding the contact sensor modalities which are latched topics, the recorded data contains 20 time-synchronized ROS topics. For the offline-matching recorder, we use all five closest samples for the calculation.

TABLE II
Dataset Distribution, User Groups defined in section IV-B

Training Task	User Group	Number of Instances		Total
		Online	Offline	
Suturing and Knot Tying	N	13	2	104
	E	36	12	
	P	2	39	
Peg Transfer	N	7	-	18
	E	11		
Tissue Manipulation	N	9	-	21
	E	12		
Dissection	N	6	-	71
	E	15	9	
	P	1	40	

For both recorders, we select the ROS image topic timestamp of the left stereo endoscope camera as the reference timestamp to perform time latency analysis for synchronization quality evaluation across different modalities. Table III summarizes descriptive statistics and Fig. 8 shows the raw time latency distribution. The few outliers could be due to: (1) the time latency for the control loop of the dVRK software, (2) extra time to numerically solve inverse kinematics near singularities, and (3) suboptimal CPU performance.

The stereo camera parameters are included in the dataset. Notably, our post-collection processing tool can also properly handle data subsets involving endoscope movements if the hand-eye calibration is performed.

TABLE III
Synchronized Recorder Comparison and Time Analysis

Attribute	Online-matching	Offline-matching
Time latency mean $\pm$ std (ms)	6.36 $\pm$ 4.72	1.35 $\pm$ 0.81
Time latency median (ms)	5.58	1.33
Recording frequency (Hz)	4.04 $\pm$ 1.69	10
Ready to use	Yes	No (post-collection time matching and interpolation required)

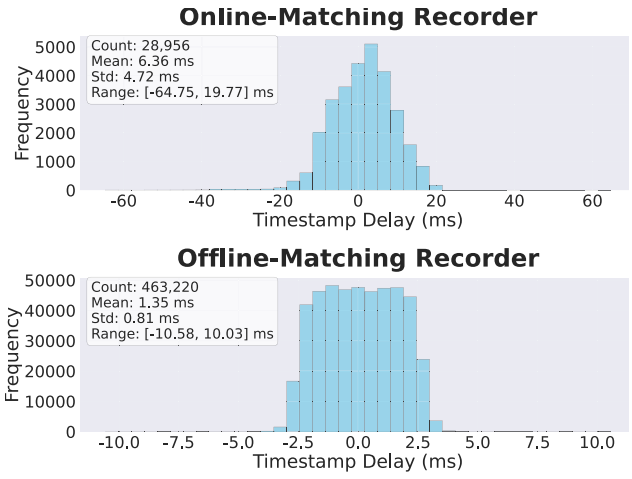

Fig. 8. Overall raw time latency distribution of all ROS topics

D. Dataset Validation in Skill Assessment

We evaluate our synchronized data on the task of skill assessment to demonstrate the feasibility of the collected data and the post-collection processing toolbox. We follow a data-driven regression approach using the unified multi-path framework for automatic surgical skill assessment [15].

After interpolating the second-stage outputs from the offline-matching recorder, the resulting dataset shares identical data structure with the online-matching one. We selected 43 instances from 8 users with varying experience levels. Each instance is split into suturing or knot-tying segments based on manual annotations.

An experienced user with knowledge of animal surgery graded all instances using the global rating score (GRS) [15] in the range of 6 to 30. The GRS is derived from an objective rubric [48] with six categories: (1) respect for tissue, (2) suture/needle handling, (3) time and motion, (4) flow of operation, (5) quality of final product, and (6) overall performance. Each category is scored on a scale of 1–5.

The model takes three kinds of synchronized input modalities: (i) kinematic features (14D), including Cartesian positions (6D), measured velocities (6D), and gripper openings (2D); (ii) visual features (2048D) extracted from a ResNet-101 encoder applied to RGB images; and (iii) one-hot encoded gesture labels (14D) based on 14 common surgical gestures, following the JIGSAWS gesture taxonomy [34]. All modalities are aligned per-frame using our synchronized recorder to ensure coherent temporal context across streams.

To ensure diverse coverage of skill levels, we construct four stratified folds for cross-validation [49]. Each fold partitions data into training and test sets with a balance of GRS distribution and user identity.

Our model follows the multi-path design of Liu et al. [15], with temporal encoders for kinematic (T), visual (V), and gesture (E) streams. Each head outputs the five GRS scores and is trained with equal-weight mean squared error on normalized labels. A temporal contrastive loss is added for regularization. We report Spearman's rank correlation coefficient (SROCC) averaged over four folds (Table IV).

TABLE IV
SKILL ASSESSMENT PERFORMANCE ON OUR SYNCHRONIZED DATASET. SROCC IS REPORTED FOR EACH CROSS-VALIDATION FOLD.

Task	Fold 1	Fold 2	Fold 3	Fold 4	Mean $\pm$ Std
Suturing	0.658	0.828	0.844	0.881	**0.803 $\pm$ 0.086**
Knot Tying	0.870	0.875	0.618	0.699	**0.765 $\pm$ 0.111**

VI. Discussion and Conclusion

In this manuscript, we propose SurgSync, a multi-modal data collection framework with (1) dual-mode (online/offline) time-synchronized recorder, (2) advanced hardware integration of a modern stereo endoscope and a novel contact sensor on the dVRK-Si, and (3) a post-collection toolbox for rectification, depth estimation and a kinematic reprojection approach using a Gaussian heatmap. We conduct a user study to acquire datasets and demonstrate feasibility by training and evaluating a skill-assessment model on the collected data. The performance of the synchronized recorders is highly dependent on the workstation hardware, yet it can still obtain reliable data even with suboptimal hardware. All open-source software and data are available at https://surgsync.github.io/

In future work, we will perform additional data collection using the offline-matching approach to build a large-scale training dataset. Moreover, we will include additional data with more types of instruments and training tasks and add experiments with in-vivo or cadaver environments if possible.

ACKNOWLEDGMENT

Thanks to Dale Bergman, Alessandro Gozzi and the Intuitive Foundation for all the hardware support of the dVRK-Si. Thanks to Cornerstone Robotics Ltd. for providing the modern chip-on-tip endoscope.

References

[1] C. D'Ettorre, A. Mariani, A. Stilli, F. R. y Baena, P. Valdastri, A. Deguet *et al.*, "Accelerating Surgical Robotics Research: A Review of 10 Years With the da Vinci Research Kit," *IEEE Robotics & Automation Magazine*, vol. 28, no. 4, pp. 56–78, 2021.

[2] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee *et al.*, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models: Open X-Embodiment Collaboration 0," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903.

[3] T. Haidegger, S. Speidel, D. Stoyanov, and R. M. Satava, "Robot-Assisted Minimally Invasive Surgery—Surgical Robotics in the Data Age," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 835–846, 2022.

[4] H. Ding, J. Zhang, P. Kazanzides, J. Y. Wu, and M. Unberath, "CaRTS: Causality-driven robot tool segmentation from vision and kinematics data," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2022, pp. 387–398.

[5] C. D'Ambrosia, F. Richter, Z.-Y. Chiu, N. Shinde, F. Liu, H. I. Christensen *et al.*, "Robust surgical tool tracking with pixel-based probabilities for projected geometric primitives," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 455–15 462.

[6] N. Fernandes, E. Oliveira, and N. F. Rodrigues, "Future perspectives of deep learning in laparoscopic tool detection, classification, and segmentation: A systematic review," in *IEEE Intl. Conf. on Serious Games and Applications for Health (SeGAH)*. IEEE, 2023, pp. 1–8.

[7] H. Ding, Y. Zhang, T. Lu, R. Liang, H. Shu, L. Seenivasan *et al.*, "SegSTRONG-C: Segmenting surgical tools robustly on non-adversarial generated corruptions–an EndoVis' 24 challenge," *arXiv preprint arXiv:2407.11906*, 2024.

[8] H. Xu, A. Weld, C. Xu, A. Roddan, J. Cartucho, M. A. Karaoglu *et al.*, "SurgRIPE challenge: Benchmark of surgical robot instrument pose estimation," *Medical Image Analysis*, p. 103674, 2025.

[9] Z. Wu, A. Schmidt, R. Moore, H. Zhou, A. Banks, P. Kazanzides *et al.*, "SurgPose: a dataset for articulated robotic surgical tool pose estimation and tracking," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.

[10] N. Yilmaz, J. Y. Wu, P. Kazanzides, and U. Tumerdem, "Neural network based inverse dynamics identification and external force estimation on the da Vinci Research Kit," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1387–1393.

[11] Z. Chua, A. M. Jarc, and A. M. Okamura, "Toward force estimation in robot-assisted surgery using deep learning with vision and robot state," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 12 335–12 341.

[12] M. D. I. Reyzabal, M. Chen, W. Huang, S. Ourselin, and H. Liu, "DaFoEs: Mixing datasets towards the generalization of vision-state deep-learning force estimation in minimally invasive robotic surgery," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2527–2534, 2024.

[13] B. Varadarajan, C. Reiley, H. Lin, S. Khudanpur, and G. Hager, "Data-derived models for segmentation with application to surgical assessment and training," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2009, pp. 426–434.

[14] A. Zia and I. Essa, "Automated surgical skill assessment in RMIS training," *Intl. Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 5, pp. 731–739, 2018.

[15] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan *et al.*, "Towards unified surgical skill assessment," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 9522–9531.

[16] K. Lam, J. Chen, Z. Wang, F. M. Iqbal, A. Darzi, B. Lo *et al.*, "Machine learning for technical skill assessment in surgery: a systematic review," *NPJ Digital Medicine*, vol. 5, no. 1, p. 24, 2022.

[17] H. Zhou, Y. Jiang, S. Gao, S. Wang, P. Kazanzides, and G. S. Fischer, "Suturing tasks automation based on skills learned from demonstrations: A simulation study," in *Intl. Symp. on Medical Robotics (ISMR)*. IEEE, 2024, pp. 1–8.

[18] P. M. Scheikl, N. Schreiber, C. Haas, N. Freymuth, G. Neumann, R. Lioutikov *et al.*, "Movement primitive diffusion: Learning gentle robotic manipulation of deformable objects," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5338–5345, 2024.

[19] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn *et al.*, "Surgical Robot Transformer (SRT): Imitation learning for surgical tasks," *arXiv preprint arXiv:2407.12998*, 2024.

[20] J. W. Kim, J.-T. Chen, P. Hansen, L. X. Shi, A. Goldenberg, S. Schmidgall *et al.*, "SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Science Robotics*, vol. 10, no. 104, p. eadt5254, 2025.

[21] Y. Long, A. Lin, D. H. C. Kwok, L. Zhang, Z. Yang, K. Shi *et al.*, "Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery," *Science Robotics*, vol. 10, no. 104, p. eadt3093, 2025.

[22] E. Colleoni, P. Edwards, and D. Stoyanov, "Synthetic and real inputs for tool segmentation in robotic surgery," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2020, pp. 700–710.

[23] T. Zeng, G. Loza Galindo, J. Hu, P. Valdastri, and D. Jones, "Realistic surgical image dataset generation based on 3D Gaussian Splatting," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2024, pp. 510–519.

[24] J. A. Barragan, J. Zhang, H. Zhou, A. Munawar, and P. Kazanzides, "Realistic Data Generation for 6D Pose Estimation of Surgical Instruments," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024, pp. 13 347–13 353.

[25] J. Wu, H. Zhou, P. Kazanzides, A. Munawar, and A. Liu, "SurgicAI: A hierarchical platform for fine-grained surgical policy learning and benchmarking," *Advances in Neural Information Processing Systems*, vol. 37, pp. 63 771–63 789, 2024.

[26] S. Yang, Z. Wu, M. Hong, Q. Li, D. Shen, S. E. Salcudean, and Y. Jin, "Instrument-Splatting: Controllable photorealistic reconstruction of surgical instruments using Gaussian Splatting," *arXiv preprint arXiv:2503.04082*, 2025.

[27] Z. Wu, A. Schmidt, P. Kazanzides, and S. E. Salcudean, "Augmenting efficient real-time surgical instrument segmentation in video with point tracking and Segment Anything," *Healthcare Technology Letters*, vol. 12, no. 1, p. e12111, 2025.

[28] M. Moghani, N. Nelson, M. Ghanem, A. Diaz-Pinto, K. Hari, M. Azizian *et al.*, "SuFIA-BC: Generating high quality demonstration data for visuomotor policy learning in surgical subtasks," *arXiv preprint arXiv:2504.14857*, 2025.

[29] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An Open-Source Research Kit for the da Vinci® Surgical System," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 6434–6439.

[30] K. Xu, J. Y. Wu, A. Deguet, and P. Kazanzides, "dVRK-Si: The Next Generation da Vinci Research Kit," in *IEEE Intl. Symp. on Medical Robotics (ISMR)*, 2025, pp. 185–191.

[31] J. Xu, B. Li, B. Lu, Y.-H. Liu, Q. Dou, and P.-A. Heng, "SurRol: An open-source reinforcement learning centered and dVRK compatible platform for surgical robot learning," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021, pp. 1821–1828.

[32] A. Munawar, J. Y. Wu, G. S. Fischer, R. H. Taylor, and P. Kazanzides, "Open simulation environment for learning and practice of robot-assisted surgical suturing," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3843–3850, 2022.

[33] Q. Yu, M. Moghani, K. Dharmarajan, V. Schorp, W. C.-H. Panitch, J. Liu *et al.*, "ORBIT-Surgical: An open-simulation framework for learning surgical augmented dexterity," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024, pp. 15 509–15 516.

[34] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin *et al.*, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, 2014, p. 3.

[35] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro *et al.*, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Trans. on Biomedical Engineering*, vol. 64, no. 9, pp. 2025–2041, 2017.

[36] A. Moglia, K. Georgiou, E. Georgiou, R. M. Satava, and A. Cuschieri, "A systematic review on artificial intelligence in robot-assisted surgery," *International Journal of Surgery*, vol. 95, p. 106151, 2021.

[37] A. Hendricks, M. Panoff, K. Xiao, Z. Wang, S. Wang, and C. Bobda, "Exploring the limitations and implications of the JIGSAWS dataset for robot-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9534–9541, 2024.

[38] Y. Sharon, T. Nevo, D. Naftalovich, L. Bahar, Y. Refaely, and I. Nisky, "Augmenting robot-assisted pattern cutting with periodic perturbations-can we make dry lab training more realistic?" *IEEE Trans. on Biomedical Engineering*, 2024.

[39] I. Rivas-Blanco, C. J. P. Del-Pulgar, A. Mariani, G. Tortora, and A. J. Reina, "A surgical dataset from the da Vinci Research Kit for task automation and recognition," in *Intl. Conf. on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2023, pp. 1–6.

[40] S. Liu, B. Zheng, W. Chen, Z. Peng, Z. Yin, J. Shao *et al.*, "A comprehensive evaluation of multi-modal large language models for endoscopy analysis," *arXiv preprint arXiv:2505.23601*, 2025.

[41] P. Badger and P. Stoffregen, "Arduino Capacitive Sensing Library," 2016. [Online]. Available: https://playground.arduino.cc/Main/CapacitiveSensor/index-2.html

[42] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, 2013.

[43] T. E. Lee, J. Tremblay, T. To, J. Cheng, T. Mosier, O. Kroemer, D. Fox, and S. Birchfield, "Camera-to-robot pose estimation from a single image," in *Intl. Conf. on Robotics and Auto. (ICRA)*. IEEE, 2020, pp. 9426–9432.

[44] B. Burkhart and A. Deguet, "dVRK hand-eye calibration package," 2022. [Online]. Available: https://github.com/jhu-dvrk/dvrk_camera_registration

[45] Z. Cui, J. Cartucho, S. Giannarou, and F. R. y Baena, "Caveats on the first-generation da Vinci Research Kit: latent technical constraints and essential calibrations," *IEEE Robotics & Automation Magazine*, 2023.

[46] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "FoundationStereo: Zero-shot stereo matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 5249–5260.

[47] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conf. on Computer Vision*. Springer, 2020, pp. 402–419.

[48] J. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison *et al.*, "Objective structured assessment of technical skill (OSATS) for surgical residents," *British Journal of Surgery*, vol. 84, no. 2, pp. 273–278, 1997.

[49] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu *et al.*, "Uncertainty-aware score distribution learning for action quality assessment," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 9839–9848.