

Multimodal AI for Automated Assessment of Surgical Performance

Shekhar Madhav Khairnar

`shekharmadhav.khairnar@utsouthwestern.edu`

UT Southwestern Medical Center <https://orcid.org/0009-0001-3670-6342>

Huu Phong Nguyen

UT Southwestern Medical Center

Sofia Garces-Palacios

UT Southwestern Medical Center

Samy Castillo

UT Southwestern Medical Center

Andres Abreu

UT Southwestern Medical Center

Amr Al Abbas

UT Southwestern Medical Center

Kaustubh Gopal

UT Southwestern Medical Center

Daniel Scott

UT Southwestern Medical Center

Herbert J Zeh III

UT Southwestern Medical Center

Patricio Polanco

UT Southwestern Medical Center

Ganesh Sankaranarayanan

`Ganesh.Sankaranarayanan@UTSouthwestern.edu`

UT Southwestern Medical Center

Research Article

Keywords: Automated skill assessment, Spatio-temporal Modeling, AI, Multimodal Fusion

Posted Date: July 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10271150/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Version of Record: A version of this preprint was published at npj Digital Surgery on August 7th, 2026.
See the published version at <https://doi.org/10.1038/s44484-026-00018-2>.

Multimodal AI for Automated Assessment of Surgical Performance

Shekhar Madhav Khairnar¹, Huu Phong Nguyen¹, Sofia Garces-Palacios¹, Samy Castillo¹, Andres Abreu¹, Amr Al Abbas¹, Kaustubh Gopal¹, Daniel Scott¹, Herbert J Zeh III¹, Patricio Polanco¹, Ganesh Sankaranarayanan^{1*}

¹Department of Surgery, UT Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas, 75390, United States

*Corresponding author email:

Ganesh.Sankaranarayanan@UTSouthwestern.edu

Contributing authors:

shekharmadhav.khairnar@utsouthwestern.edu;

huuphong.nguyen@utsouthwestern.edu;

Diana.GarcesPalacios@UTSouthwestern.edu;

Samy.SamyCastillo@UTSouthwestern.edu;

Andres.Abreu@UTSouthwestern.edu;

Amr.AlAbbas@UTSouthwestern.edu;

Kaustubh.Gopal@UTSouthwestern.edu;

Daniel.Scott@UTSouthwestern.edu;

Herbert.Zeh@UTSouthwestern.edu;

Patricio.Polanco@UTSouthwestern.edu;

Abstract

Accurate and objective assessment of surgical skill is crucial for improving training outcomes, ensuring patient safety, and enabling scalable evaluation across clinical and simulation contexts. Current approaches often depend on subjective human ratings or unimodal analysis, limiting reliability and applicability in real-world settings. We propose a novel multimodal deep learning framework that integrates visual and kinematic data to evaluate surgical proficiency automatically. Our method leverages YOLOv8 and BYTE Tracker for surgical tool tracking and trajectory extraction, an autoencoder for encoding tool motion, ResNet-50 for deriving semantic features from heatmap density plots, and a SlowFast-R50 backbone for capturing spatio-temporal video cues. These latent features are fused and input to a 1D convolutional neural network to predict expert-assigned skill scores. On a JIGSAWS simulation benchmark (consisting of 103 videos), our model achieved a Spearman correlation of **0.85** in 4-fold cross-validation and **0.87** in a leave-one-user-out evaluation, which outperforms prior approaches. On a Biotissue dataset (205 videos including GJ, PJ, HJ), fusing kinematic and visual features yielded a top Spearman correlation of **0.67** (GJ dataset) and **0.69** (PJ dataset). These results demonstrate our framework's scalability, robustness, and potential for clinical deployment.

Keywords: Automated skill assessment, Spatio-temporal Modeling, AI, Multimodal Fusion

Introduction

Accurate and objective assessment of surgical skill is essential for enhancing patient safety, improving surgical education, and ensuring consistent clinical outcomes. Historically, the evaluation of technical skills has relied on frameworks such as the Objective Structured Assessment of Technical Skills (OSATS)¹ and Global Rating Scale (GRS)². While these measures introduced rigor and structure into surgical assessment, they still rely on subjective expert ratings, which can be affected by inter-rater variability, limited scalability, and significant resource requirements³.

Recent advances in artificial intelligence (AI), computer vision, and deep learning have accelerated the development of automated methods for surgical skill assessment. Leveraging the availability of synchronized kinematic and video data, research efforts such as the Johns Hopkins University and Intuitive Surgical, Inc., Gesture and Skill Assessment Working Set (JIGSAWS)⁴ have enabled the design and benchmarking of machine learning models for skill evaluation. Deep neural networks have demonstrated their ability to model complex spatial and temporal patterns inherent in surgical tasks, leading to promising results in both gesture recognition and skill classification using Temporal Convolutional Network (TCN) and EndoNet models^{5,6}. In earlier studies, researchers explored frequency-based video analysis for skill assessment⁷ and automated assessment systems tailored for surgical training environments⁸.

However, a significant limitation of many existing approaches is their reliance on either a single data modality (e.g., only video or only kinematics) or on precise sensor telemetry available primarily in robotic-assisted surgeries⁹. This restricts their applicability in real-world clinical settings, where such comprehensive sensor data may not be available and where data variability is high due to factors like camera motion and lighting changes. To address these challenges, recent methods have employed multimodal fusion strategies integrating visual, kinematic, and event data for more robust and generalizable skill assessment^{9,10}.

Recently, contrastive learning and transformer-based architectures have emerged as highly effective means for automated surgical skill assessment. Unified video-based frameworks, such as those proposed by Liu et al.¹⁰ and Yanik et al.¹¹, demonstrate strong performance for both formative and summative assessment by modeling complex spatio-temporal dependencies directly from surgical videos. Building on this foundation, methods such as group-aware contrast regression (CoRe + GART)¹² introduce exemplar-based learning, where models are trained to regress relative skill scores between a candidate video and a set of reference exemplars using group-aware regression trees for hierarchical supervision. The Contra-Sformer

framework¹³ refines this approach by utilizing reference-based multi-head attention, contrasting each trial with a “gold standard” expert demonstration and leveraging a regression transformer to focus on fine-grained, inter-video skill differences.

Current models with state-of-the-art performance on the JIGSAWS benchmark, as evaluated using the standard 4-fold cross-validation split, include RICA2¹⁴, which achieves an average Spearman rank correlation coefficient (SCC) of 0.92 through a rubric-informed, graph-based model with calibrated uncertainty estimation. Similarly, Temporal Parsing Transformer (TPT)¹⁵ utilizes a transformer-based design to obtain a strong average correlation of 0.89. Purely video-based approaches including CoRe + GART¹² and Contra-Sformer¹³ demonstrate that advanced contrastive regression and attention mechanisms can achieve high predictive accuracy (average correlations of 0.85 and 0.81, respectively) without the need for explicit kinematic or event sequence inputs. Additional models such as MultiPath-VTPE¹⁰ and Video Semantic Aggregation (ViSA)¹⁶ extend this trend, further validating the potential of sophisticated visual and temporal modeling techniques, as they consistently report average correlations exceeding 0.80. Collectively, these advances illustrate both the rapid progress and the increasing methodological diversity of deep learning approaches in the automated assessment of surgical skill. A comprehensive review of automated skill assessment in robotic minimally invasive surgery has been provided by Zia et al. and Youssef et al.^{17,18}. Despite these advances, an important practical gap remains. Existing methods often achieve strong performance either by relying primarily on a single dominant modality or by depending on specialized inputs and modeling assumptions that may not be consistently available across datasets and assessment settings. In contrast, our framework is designed to support skill assessment under availability of variable modality by integrating complementary signals when present and by deriving motion-based representations directly from video through tool segmentation, tracking, and trajectory extraction when direct kinematic measurements are unavailable. This makes the framework adaptable across benchmark and simulation-based settings with differing annotation and sensing constraints.

In this work, we introduce a novel multimodal deep learning framework for automated surgical skill assessment, designed to address the limitations of prior approaches. Our method integrates multiple complementary feature channels, including spatio-temporal video features (V), kinematic motion features (K), and semantic density features (D) extracted from all surgical videos. For the JIGSAWS dataset, we further incorporate event sequence features (E) and task completion time (P) to leverage its richer annotation. We rigorously evaluate our framework on two complementary benchmarks: the JIGSAWS dataset, which offers controlled,

well-annotated simulation data, and the Biotissue dataset, which represents greater clinical variability and lacks direct sensor-based kinematics as well as event and task completion time annotations. The purpose of this evaluation is to systematically assess the generalizability, robustness, and practical effectiveness of our approach across both standardized simulation and real-world clinical scenarios. Strong performance using both datasets demonstrates the scalability, objectivity, and adaptability of our framework for surgical skill assessment in both research and clinical contexts. Although the Biotissue dataset is simulation-based rather than clinical, it represents a high-fidelity, procedure-specific training environment within our validated robotic curriculum. Our prior work^{19,20} has shown that performance on the simulated gastrojejunostomy and pancreatojejunostomy task is a significant predictor of intraoperative technical performance on the same task in the operating room, supporting the translational relevance of this simulation setting. However, the present study does not by itself establish generalizability to real surgical cases. In addition, extension to laparoscopic settings would require further validation and adaptation to differences in camera configuration, instrument appearance, motion characteristics, and tracking conditions.

The primary contributions of this work are:

- A comprehensive multimodal feature extraction pipeline that unifies spatio-temporal video features (**V**), kinematic features (**K**), semantic density features (**D**), and where available, event sequence (**E**) and task completion time (**P**) information for surgical skill assessment.
- A robust and generalizable framework validated on both simulated (JIGSAWS: **V**, **K**, **D**, **E**, **P**) and real-world (Biotissue: **V**, **K**, **D**) surgical datasets.
- Extensive experimental evidence demonstrates the effectiveness of our method in challenging and diverse environments, with systematic ablation across all feature channels.

Results

Segmentation Performance with YOLOv8

We fine-tuned YOLOv8, pretrained on COCO, on our Biotissue videos to segment needle drivers into wrist, shaft, and jaw parts, labeling each instance as left or right. Initial training on 1239 manually annotated frames yielded a segmentation performance of $\text{mAP}@50-95 = 0.57$. By augmenting the training set with 152 additional masks automatically generated using SAM/SAM2 (with YOLOv8 bounding boxes as prompts), we increased the total to 1391 training images and improved $\text{mAP}@50-95$ to 0.61. Although the gain in segmentation accuracy was modest, the SAM2-assisted workflow substantially reduced the burden of manual pixel-level annotation, which is labor-intensive and difficult to scale for long surgical videos,

while still enabling efficient expansion of the training set with sufficiently reliable masks for downstream trajectory analysis. These results demonstrate that high-quality automated mask generation with SAM2 can efficiently expand the annotated dataset and measurably improve segmentation performance.

Table 1: Comparison of methods on the JIGSAWS dataset (4-fold). Abbreviations for inputs: E, event sequence; K, kinematic features; P, task completion time; V, video features. Other abbreviations: 1D-CNN, 1-dimensional convolutional neural network; TPT, temporal parsing transformer. Values are reported as mean $\pm$ standard deviation.

Method	Input features	Spearman’s correlation coefficient $\uparrow$			
		SU	NP	KT	Avg
Our 1D-CNN	V, K, E	0.75 ± 0.14	0.87 ± 0.08	0.90 ± 0.04	0.85
CoRe + GART ¹²	V	0.84	0.86	0.86	0.85
ViSA ¹⁶	V	0.79	0.86	0.84	0.83
Contra-Sformer ¹³	V	0.69 ± 0.15	0.81 ± 0.11	0.87 ± 0.04	0.81
RICA2 ¹⁴	V	0.92	0.94	0.90	0.92
TPT ¹⁵	V	0.88	0.88	0.91	0.89
MultiPath (VTPE) ¹⁰	V, K, E, P	0.82	0.76	0.83	0.81
Mean absolute error $\downarrow$					
Our 1D-CNN	V, K, E	2.88 ± 0.58	2.00 ± 0.65	2.43 ± 0.54	2.44
Contra-Sformer ¹³	V	2.99 ± 0.75	3.21 ± 1.12	2.10 ± 0.47	2.77
ViSA ¹⁶	V	2.70	2.05	2.60	2.45

JIGSAWS Dataset (4-Fold Cross-Validation and LOUO)

Performance compared with state-of-the-art approaches is summarized in Table 1, which presents the results of recent methods on the JIGSAWS dataset with 4-fold cross-validation, including both SCC and MAE where available. Methods included in Tables 1 and 2 were selected based on the evaluation settings and metrics reported in the respective source papers. Table 1 summarizes 4-fold cross-validation results, whereas Table 2 summarizes leave-one-user-out (LOUO) results. Because not all prior studies reported results for both protocols, and not all studies reported both Spearman’s correlation coefficient and mean absolute error, the set of comparison methods differs across tables and metrics. Accordingly, absence of a method

from a given table or metric column reflects lack of directly reported results in the original publication rather than exclusion from comparison. Notably, ViSA and Contra-Sformer report MAE as well as SCC, so to ensure a fair and comprehensive comparison, we report both these metrics in our experiments. In the multimodal setting, our 1D-CNN achieved the highest average SCC (0.85) among all methods leveraging multiple input modalities (kinematics, video, event, and optionally density or completion time), highlighting the benefits of combining complementary information streams. For video-only models, RICA2 attains the highest average SCC (0.92), attributable not to a larger or more complex vision backbone, but to its novel probabilistic framework that incorporates score rubrics and models prediction uncertainty using a graph neural network with stochastic embeddings¹⁴. This design leads to both accurate score predictions and calibrated uncertainty estimates. Furthermore, our approach delivered the lowest MAE (2.44) in absolute terms compared to previous state-of-the-art approaches, confirming the strong predictive performance of our multimodal model. Together, these findings reinforce the importance of modality fusion for robust surgical skill assessment and highlight the innovation of RICA2 among video-based approaches.

Table 2: Comparison of methods on JIGSAWS (Leave-one-user-out (LOUO)) (Spearman’s correlation coefficients). Abbreviations for inputs: E, event sequence; K, kinematic features; P, task completion time; V, video features. Other abbreviations: 1D-CNN, 1-dimensional convolutional neural network; KT, knot-tying; NP, needle-passing; SU, suturing. Values are reported as mean $\pm$ standard deviation.

Method	Input Features	Spearman’s correlation coefficient $\uparrow$			
		SU	NP	KT	Avg
Our 1D-CNN	K, E	0.74 $\pm$ 0.34	0.70 $\pm$ 0.29	0.94 $\pm$ 0.07	0.83
Our 1D-CNN	V	0.91 $\pm$ 0.10	0.86 $\pm$ 0.18	0.80 $\pm$ 0.26	0.87
ViSA ¹⁶	V	0.76	0.90	0.72	0.81
Contra-Sformer ¹³	V	0.65 $\pm$ 0.14	0.71 $\pm$ 0.30	0.69 $\pm$ 0.19	0.68 $\pm$ 0.22
MTL-VF (C3D) ²¹	V	0.69	0.86	0.83	0.80
VBA NET ¹¹	K	0.49	0.75	0.83	0.72
MTL-VF (ResNet) ²¹	V	0.68	0.48	0.72	0.64
MultiPath (VTPE) ¹⁰	V, K, P, E	0.45	0.65	0.59	0.57
MultiPath (TP) ¹⁰	K, P	0.40	0.63	0.55	0.53

MultiPath (VTP) ¹⁰	V, K, P	0.45	0.52	0.60	0.53
JRG 39 ⁴⁴	V, K	0.35	0.67	0.19	0.40
AIM 17 ⁴⁵	V, K	0.45	0.34	0.61	0.47

In the LOUO evaluation (Table 2), our model demonstrated strong generalization to unseen surgeons by leveraging different combinations of input features. Kinematic and event sequence characteristics (K, E) alone yielded a strong average SCC of 0.83, with particularly high performance on the knot-tying task (KT; 0.94 ± 0.07), outperforming all previous methods on this subtask. Using only visual features (V), our 1D-CNN achieved the best average correlation overall (0.87), with outstanding results for suturing (SU; 0.91 ± 0.10) and needle-passing (NP; 0.86 ± 0.18). Previous state-of-the-art approaches, such as ViSA¹⁶ and MTL-VF (C3D, ResNet)²¹, are reported to have lower average correlations (ranging from 0.64 to 0.81), and generally do not match the highest scores achieved by our visual-only or kinematic and event models.

Table 3: Four-fold ablation study: Spearman’s correlation coefficients on the JIGSAWS dataset. Abbreviations for inputs: D, semantic density features; E, event sequence; K, kinematic features; P, task completion time; V, video features. Other abbreviations: KT, knot-tying; NP, needle-passing; SU, suturing. Values are reported as mean $\pm$ standard deviation.

Input features	Spearman’s correlation coefficient $\uparrow$			
	SU	NP	KT	Avg
V, K, E, D, P	0.81 ± 0.07	0.82 ± 0.15	0.82 ± 0.06	0.82
V, K, E, D	0.79 ± 0.06	0.79 ± 0.18	0.79 ± 0.18	0.76
V, K, E	0.75 ± 0.14	0.87 ± 0.08	0.90 ± 0.04	0.85
K, E	0.79 ± 0.01	0.51 ± 0.61	0.71 ± 0.18	0.69
P	0.39 ± 0.12	0.20 ± 0.57	0.47 ± 0.28	0.36
V	0.62 ± 0.24	0.93 ± 0.05	0.75 ± 0.07	0.80

In 4-fold cross-validation (Table 3), our multimodal approach combining kinematic (K), video (V), and event sequence (E) features achieved an average SCC of 0.85 with GRS, outperforming strong multimodal baselines such as that of MultiPath¹⁰.

We additionally compared the proposed 1D-CNN with alternative regression heads using the same PCA-compressed V+K+E feature representation, as this configuration yielded the strongest multimodal performance. The results are summarized in Table 4.

Table 4: Comparison of alternative regression heads on the JIGSAWS dataset using the PCA-compressed V+K+E feature representation under 4-fold cross-validation. Abbreviations for inputs: E, event sequence; K, kinematic features; V, video features. Task abbreviations: KT, knot-tying; NP, needle-passing; SU, suturing. Values are reported as mean $\pm$ standard deviation.

Method	Input features	Spearman’s correlation coefficient $\uparrow$			
		SU	NP	KT	Avg
Linear Regression		0.67 ± 0.11	0.48 ± 0.40	0.68 ± 0.20	0.61
Ridge		0.67 ± 0.11	0.48 ± 0.40	0.68 ± 0.20	0.61
SVR	V, K, E	0.73 ± 0.05	0.49 ± 0.57	0.67 ± 0.08	0.63
Random Forest		0.40 ± 0.29	0.19 ± 0.24	0.48 ± 0.13	0.36
MLP		0.31 ± 0.49	-0.06 ± 0.60	0.61 ± 0.18	0.29

Table 3 presents our detailed ablation study evaluating the effect of different combinations of input characteristics on the SCC in all JIGSAWS tasks. We observed that multimodal integration of video, kinematics, event sequences, density, and task completion time (V, K, E, D, P) achieved the highest SCC for suturing (0.81 ± 0.07), demonstrating the value of modality fusion for capturing diverse aspects of surgical skill. In particular, removing density and completion time (V, K, E) maintained strong average performance (0.85), with the highest knot-tying SCC (0.90 ± 0.04), while kinematics alone (K, E) and video alone (V) resulted in lower mean correlations and increased variance.

Because task completion time (P) may act as a shortcut variable, we additionally evaluated a P-only baseline in JIGSAWS to assess its independent predictive value. The P-only model achieved an average SCC of 0.36, which was substantially lower than the video-only baseline (0.80) and the best-performing multimodal configuration V+K+E (0.85). Moreover, adding P did not improve performance over V+K+E alone, as the average SCC changed from 0.85 for V+K+E to 0.82 for V+K+E+D+P. These results suggest that task completion time alone is not a reliable proxy for technical skill in our setting and that the strongest predictive performance

does not depend on P. Accordingly, P should be interpreted as a benchmark-specific auxiliary variable rather than a required practical input for skill assessment.

Notably, the addition of semantic density features (D) to the (V, K, E) configuration resulted in a performance decrease (average SCC of 0.76). We attribute this drop to the nature of the density plots in the JIGSAWS context; because these plots are generated from 2D Cartesian (x , y) coordinates that have been normalized to a $[0, 1]$ range rather than representing absolute 3D workspace trajectories, they may introduce significant spatial noise into the semantic representation. However, the model’s performance recovered to 0.82 upon the inclusion of task completion time (P). This suggests that completion time serves as a powerful global indicator of efficiency, providing a high-signal-to-noise ratio that partially offsets the coordinate-based noise introduced by the density features.

To justify the choice of the 1D-CNN regression head, we performed an additional comparison using the V+K+E configuration, which yielded the strongest multimodal performance in the JIGSAWS ablation study. Specifically, we compared the proposed 1D-CNN with several simpler models trained on the same PCA-compressed V+K+E feature representation under the 4-fold cross-validation setting, including linear regression, ridge regression, SVR with an RBF kernel, random forest, and MLP (Table 4). Among these baselines, SVR achieved the highest overall mean Spearman rank correlation (0.63), followed by linear regression and ridge regression (both 0.61), whereas random forest and MLP yielded lower mean correlations (0.36 and 0.29, respectively). In comparison, the proposed 1D-CNN achieved a substantially higher mean Spearman correlation of 0.85 on the same V+K+E configuration. These results indicate that the observed performance gain is not attributable to PCA-based dimensionality reduction alone, but also to the capacity of the 1D-CNN to model nonlinear interactions within the compressed multimodal feature space.

Table 5: Leave-one-user-out (LOUO) evaluation and ablation study on the JIGSAWS dataset. Abbreviations for inputs: D, semantic density features; E, event sequence; K, kinematic features; P, task completion time; V, video features. Other abbreviations: 1D-CNN, 1-dimensional convolutional neural network; KT, knot-tying; NP, needle-passing; SU, suturing. Values are reported as mean $\pm$ standard deviation.

Method	Input features	Spearman’s correlation coefficient $\uparrow$			
		SU	NP	KT	Avg
Our 1D-CNN	V, K, E, D, P	0.56 ± 0.40	0.55 ± 0.51	0.78 ± 0.25	0.64
Our 1D-CNN	V, K, E, D	0.78 ± 0.11	0.62 ± 0.30	0.62 ± 0.25	0.68
Our 1D-CNN	V, K, E	0.76 ± 0.26	0.81 ± 0.2	0.73 ± 0.41	0.77
Our 1D-CNN	K, E	0.74 ± 0.34	0.70 ± 0.29	0.94 ± 0.07	0.83
Our 1D-CNN	V	0.91 ± 0.10	0.86 ± 0.18	0.80 ± 0.26	0.87

In the LOUO evaluation on JIGSAWS (Table 5), our multimodal models that combined kinematic and event features (K, E) generally achieved lower SCCs than video-only models, due to the high user-specific variability in kinematic and event data. These modalities tend to capture individual patterns that do not generalize well to unseen surgeons, thus limiting predictive accuracy. In contrast, video-based features are more invariant to individual style and provide better generalization, as reflected in the video-only model’s superior performance (average SCC: 0.87). Despite this challenge, our multimodal approaches, including both the K + E (average SCC: 0.83) and the full (V, K, E, D, P) configuration (average SCC: 0.64) outperformed previous multimodal baselines, with all of our variants exceeding the best MultiPath (VTPE) result (average SCC: 0.57). This trend is consistent across both LOUO and standard 4-fold cross-validation settings, as shown in Table 2 and Table 3. These results demonstrate that our framework establishes a new benchmark among multimodal models, even as video-only methods remain optimal for generalization to unseen users.

Biotissue Dataset

As shown in Table 6, our baseline model that uses only kinematic features (K), encoded with an autoencoder and RMS kinematic statistics, achieved a mean SCC of $0.47 (\pm 0.08)$ with expert OSATS scores and an MAE of $3.13 (\pm 0.46)$. Incorporating video features (V) extracted using the SlowFast-R50 model led to a substantial improvement, with the combined kinematics and video (K + V) model achieving the highest mean SCC (**0.67 ± 0.05**) and the lowest MAE (**2.71 ± 0.43**) across all settings.

Table 6: Performance on the GJ Biotissue dataset (5-fold cross-validation). Abbreviations for inputs: D, semantic density features; K, kinematic features; V, video features. Other abbreviations: MAE, mean absolute error. Values are reported as mean $\pm$ standard deviation.

Input features	Spearman correlation $\uparrow$	MAE $\downarrow$
Only K	0.47 (± 0.08)	3.13 (± 0.46)
Only D	0.36 (± 0.12)	3.38 (± 0.61)
Only V	0.57 (± 0.08)	2.88 (± 0.57)
V, D	0.61 (± 0.09)	2.85 (± 0.24)
K, D	0.54 (± 0.05)	2.98 (± 0.53)
K, V	0.67 (± 0.05)	2.71 (± 0.43)
K, V, D	0.61 (± 0.08)	2.92 (± 0.50)

Adding heatmap density features (D) to form the K + V + D configuration further improved performance relative to kinematics alone, with an SCC of 0.61 (± 0.08) and an MAE of 2.92 (± 0.50). These results suggest that semantic context from tool motion density plots provides additional useful information; however, the presence of camera movement and lack of depth estimation in the Biotissue dataset can introduce noise, limiting the full potential of the density-based features. We hypothesize that the limited utility of density heatmaps in Biotissue is partly due to camera motion, variable zoom, and the lack of depth information in the 2D representation. However, camera motion effects were not explicitly quantified, and stabilization or registration baselines were not evaluated in the present study. Overall, these findings highlight that integrating temporal kinematic data with visual and semantic information delivers a significant boost in performance for automated surgical skill assessment compared to using kinematics alone.

Table 7: Performance on the PJ Biotissue dataset (stratified 5-fold cross-validation). Abbreviations for inputs: K, kinematic features; V, video features. Other abbreviations: MAE, mean absolute error. Values are reported as mean $\pm$ standard deviation.

Input features	Spearman correlation $\uparrow$	MAE $\downarrow$
Only K	0.63 (± 0.09)	2.29 (± 0.45)
Only V	0.66 (± 0.06)	2.17 (± 0.45)
K, V	0.69 (± 0.07)	2.05 (± 0.28)

We further evaluated the framework on the pancreatojejunostomy (PJ) subset of the Biotissue dataset. As shown in Table 7, the video-only model (V) achieved a mean SCC of 0.66 with an MAE of 2.17, while the kinematic-only model (K), encoded using the autoencoder and RMS kinematic statistics, achieved a mean SCC of 0.63 with an MAE of 2.29. The combined video and kinematic model (K + V) achieved the best overall performance, with a mean SCC of 0.69 and the lowest MAE of 2.05. These findings are consistent with the GJ results and suggest that integrating visual and motion-based information provides a more effective representation for automated skill assessment than either modality alone. Density-based features were not included in the PJ experiments because they did not provide consistent benefit in the GJ setting, and their utility is likely limited in the current 2D representation by camera motion and lack of depth information.

To further evaluate the choice of kinematic trajectory representation, we compared the proposed trajectory-to-image autoencoder with direct sequence encoder baselines based on bidirectional long short term memory (BiLSTM)^{22,23}, TCN, and Transformer architectures under both K+V and K-only settings. Each sequence encoder was trained within an autoencoder framework to reconstruct the same video-derived tool trajectory inputs, and the learned latent representation from the encoder was used as the kinematic trajectory feature. Because the K+V configuration achieved the highest Spearman correlation in both the GJ and PJ Biotissue experiments, we used this setting to evaluate the effect of different kinematic trajectory encoders while keeping the video feature branch and downstream prediction model unchanged. Kinematic-only comparisons were also performed to isolate the effect of the trajectory encoder without the video feature branch. As in the original Biotissue experiments, K included the learned trajectory representation together with handcrafted motion statistics, including RMS velocity, RMS acceleration, RMS jerk, total path length.

Table 8: Comparison of trajectory sequence encoders on the GJ Biotissue dataset using the best-performing K+V feature setting. K denotes video-derived kinematic features, including

trajectory representations encoded by the specified sequence encoder and additional motion statistics such as RMS velocity, RMS acceleration, RMS jerk, displacement, and total path length. V denotes spatio-temporal video features extracted from surgical video. Abbreviations: BiLSTM, bidirectional long short-term memory; TCN, temporal convolutional network; MAE, mean absolute error. Values are reported as mean $\pm$ standard deviation across 5-fold cross-validation.

Sequence encoder for K	Input features	Spearman correlation $\uparrow$	MAE $\downarrow$
Trajectory-to-image autoencoder		0.67 (± 0.05)	2.71 (± 0.43)
BiLSTM sequence encoder	K, V	0.56 (± 0.07)	2.99 (± 0.48)
TCN sequence encoder		0.59 (± 0.06)	2.92 (± 0.41)
Transformer sequence encoder		0.57 (± 0.15)	2.99 (± 0.33)

For the GJ dataset, the trajectory-to-image autoencoder achieved the highest K+V performance, with a Spearman correlation of 0.67 ± 0.05 and MAE of 2.71 ± 0.43 (Table 8). The direct sequence encoders achieved lower correlations under the same K+V setting, including 0.56 ± 0.07 for BiLSTM, 0.59 ± 0.06 for TCN, and 0.57 ± 0.15 for Transformer. In the kinematic-only setting, the BiLSTM sequence encoder achieved a slightly higher Spearman correlation than the trajectory-to-image autoencoder, whereas the trajectory-to-image autoencoder achieved the lowest MAE (Table 9). The TCN and Transformer encoders showed lower overall performance. Although BiLSTM achieved a slightly higher kinematic-only correlation on GJ, the trajectory-to-image autoencoder remained the strongest encoder in the best-performing K+V configuration, which is the setting used in the final framework.

Table 9: Comparison of trajectory sequence encoders on the GJ Biotissue dataset using kinematic features only. K denotes video-derived kinematic features, including trajectory representations encoded by the specified sequence encoder and additional motion statistics such as RMS velocity, RMS acceleration, RMS jerk, displacement, and total path length. Abbreviations: BiLSTM, bidirectional long short-term memory; TCN, temporal convolutional network; MAE, mean absolute error. Values are reported as mean $\pm$ standard deviation across 5-fold cross-validation.

Sequence encoder for K	Input features	Spearman correlation $\uparrow$	MAE $\downarrow$
Trajectory-to-image autoencoder	K	0.47 (± 0.08)	3.13 (± 0.46)
BiLSTM sequence encoder		0.48 (± 0.11)	3.14 (± 0.63)
TCN sequence encoder		0.43 (± 0.15)	3.22 (± 0.75)
Transformer sequence encoder		0.38 (± 0.21)	3.26 (± 0.48)

For the PJ dataset, the trajectory-to-image autoencoder also achieved the highest K+V performance, with a Spearman correlation of 0.69 ± 0.07 and MAE of 2.05 ± 0.28 (Table 10). The BiLSTM, TCN, and Transformer sequence encoders achieved lower correlations of 0.55 ± 0.15 , 0.50 ± 0.10 , and 0.56 ± 0.12 , respectively. In the kinematic-only setting, the trajectory-to-image autoencoder achieved the highest Spearman correlation and lowest MAE compared with all direct sequence encoders (Table 11). These results support the use of the trajectory-to-image autoencoder as the primary kinematic representation module in the proposed framework.

Table 10: Comparison of trajectory sequence encoders on the PJ Biotissue dataset using the best-performing K+V feature setting. K denotes video-derived kinematic features, including trajectory representations encoded by the specified sequence encoder and additional motion statistics such as RMS velocity, RMS acceleration, RMS jerk, displacement, and total path length. V denotes spatio-temporal video features extracted from surgical video. Abbreviations: BiLSTM, bidirectional long short-term memory; TCN, temporal convolutional network; MAE, mean absolute error. Values are reported as mean $\pm$ standard deviation across stratified 5-fold cross-validation.

Sequence encoder for K	Input features	Spearman correlation $\uparrow$	MAE $\downarrow$
Trajectory-to-image autoencoder	K, V	0.69 (± 0.07)	2.05 (± 0.28)
BiLSTM sequence encoder		0.55 (± 0.15)	2.36 (± 0.47)
TCN sequence encoder		0.50 (± 0.10)	2.40 (± 0.31)
Transformer sequence encoder		0.56 (± 0.12)	2.44 (± 0.28)

Table 11: Comparison of trajectory sequence encoders on the PJ Biotissue dataset using kinematic features only. K denotes video-derived kinematic features, including trajectory representations encoded by the specified sequence encoder and additional motion statistics such as RMS velocity, RMS acceleration, RMS jerk, displacement, and total path length. Abbreviations: BiLSTM, bidirectional long short-term memory; TCN, temporal convolutional network; MAE, mean absolute error. Values are reported as mean $\pm$ standard deviation across stratified 5-fold cross-validation.

Sequence encoder for K	Input features	Spearman correlation $\uparrow$	MAE $\downarrow$
Trajectory-to-image autoencoder		0.63 (± 0.09)	2.29 (± 0.45)
BiLSTM sequence encoder	K	0.55 (± 0.14)	2.33 (± 0.43)
TCN sequence encoder		0.52 (± 0.10)	2.39 (± 0.34)
Transformer sequence encoder		0.57 (± 0.13)	2.40 (± 0.34)

Discussion

This work presents a unified multimodal deep learning framework for surgical skill assessment, rigorously benchmarked on both the Biotissue and JIGSAWS datasets. Our model incorporates kinematic sequences, video-based spatio-temporal features, and gesture-level event representations (for JIGSAWS), and it leverages a 1D convolutional backbone for temporal fusion. Extensive ablation and benchmarking demonstrate that integrating kinematic, video, and event cues can provide strong performance when complementary signals are available, particularly in 4-fold JIGSAWS and Biotissue evaluations, whereas video-only features yielded the best generalization in JIGSAWS leave-one-user-out evaluation.

On JIGSAWS, our best configuration of kinematics + video + event (i.e., K, V, E) achieved an average SCC of 0.85 (± 0.12) in 4-fold cross-validation (see Table 1), matching or exceeding strong recent video-based baselines such as CoRe + GART (0.85), ViSA (0.83), and ContraSformer (0.81). Notably, our method also outperformed the highly specialized MultiPath (V, T, P, E) baseline (0.81), which requires tool usage, visual, and completion time information. While RICA2 represents the strongest video-only baseline on JIGSAWS, achieving an average SCC of 0.92, this performance is enabled by a substantially more complex modeling framework that integrates probabilistic graph reasoning and rubric-driven supervision. Similarly, TPT demonstrates strong video-only performance on JIGSAWS, but it relies on temporal part

parsing and exemplar-based regression, reflecting a more specialized design than the simpler fusion-based framework used in our study. In contrast, our approach adopts a more streamlined architectural design and explicitly preserves gesture event sequences, allowing it to model execution dynamics and temporal skill patterns that extend beyond rubric alignment. Notably, despite these design differences, our method attains competitive predictive performance with lower absolute error, suggesting that explicit temporal modeling and multimodal fusion can offer an effective and more broadly applicable alternative for surgical skill assessment, particularly in settings with limited data or constrained computational resources.

In the more challenging LOUO setting an essential test of generalization to unseen surgeons our model achieved an average SCC of 0.87 when using video features alone, outperforming all other published methods, including ViSA, MTL-VF, and MultiPath, as well as recently published data for Contra-Sformer. Our approach achieved per-task peaks such as 0.91 (± 0.10) for suturing, 0.86 (± 0.18) for needle-passing, and 0.80 (± 0.26) for knot-tying. Even with multimodal input (kinematics + events), the model matches or surpasses all prior multimodal baselines on LOUO, with an SCC of 0.83.

These findings indicate that multimodal fusion should not be interpreted as universally superior to video-only modeling; rather, its value lies in exploiting complementary signals when available, particularly in settings where video alone may not capture all relevant aspects of technical performance.

The sequence encoder comparison further supports the trajectory-to-image autoencoder as the kinematic representation module. Although BiLSTM, TCN, and Transformer encoders can model temporal trajectories directly, they did not improve performance over the trajectory-to-image autoencoder in the best-performing K+V setting for either GJ or PJ. This may reflect the limited sample size and high variability of video-derived tool trajectories, where compact trajectory-image representations provide a more stable summary of global motion patterns than direct sequence models.

Ablation studies confirm that both kinematic and video inputs are crucial; neither modality alone achieves the same level of predictive power as their combination. Notably, while adding more modalities such as density heatmaps (D) or task completion time (P) can sometimes improve performance marginally, the combination of K, V, and E remains the most reliable across tasks and folds, highlighting the benefits of complementary representations.

Compared to recent pure vision-based models such as CoRe + GART, ViSA, and Contra-Sformer, our approach is competitive or superior in most evaluation protocols, while also offering greater flexibility for multimodal fusion. Unlike these other methods, which typically

require large-scale video pretraining and sophisticated backbones, our model maintains simplicity and interpretability, with lower computational overhead and easier integration into surgical robotics platforms. For Biotissue data, where direct robotic kinematic streams were not available for integration into the present assessment pipeline, our approach derives motion representations from video through tool tracking (YOLOv8 + BYTE Tracker), supporting its applicability in settings where motion cues must be recovered from video rather than provided as synchronized telemetry.

Despite these advances, several limitations remain. In the Biotissue dataset, the lack of gesture annotations restricts access to semantic cues that are available in JIGSAWS, and video-based tracking is still vulnerable to camera motion, occlusion, and calibration drift. The need for careful temporal alignment across modalities is another technical challenge. Finally, like all methods in the field, our results are limited by the size and diversity of available datasets. Future research should focus on explicit temporal synchronization, robustness to visual variability (camera shifts, zoom, and panning), and integrating pseudo-3D spatial reasoning (e.g., monocular depth estimation) to further close the gap with sensor-based systems. We also envision leveraging transformer-based temporal fusion architectures to capture long-range dependencies and doing tests on larger, more heterogeneous datasets to establish clinical utility.

This study establishes a robust and generalizable multimodal deep learning framework for automated surgical skill assessment, setting a new standard on both Biotissue and JIGSAWS datasets. Our approach combines video-derived spatio-temporal features, kinematic motion embeddings, and when available, gesture or event information, enabling flexible deployment in both sensor-rich and video-only surgical environments.

Our model achieved an average SCC of 0.87 in LOUO on JIGSAWS using only video features, surpassing all previously reported video-only methods, including ViSA¹⁶, and ContraSformer¹³. In 4-fold cross-validation, our fusion of kinematic, video, and event features (K, V, E) achieved an SCC of 0.85, representing the highest performance among multimodal approaches. While certain video-only models attain even higher correlations, our method demonstrates that integrating multiple modalities delivers state-of-the-art results for multimodal skill assessment, with reduced reliance on domain-specific feature engineering. For a Biotissue dataset, our framework enables automated skill assessment using only tool trajectories extracted from standard 2D endoscopic video, eliminating reliance on robotic telemetry and facilitating broad clinical translation.

Our future work will address challenges in temporal alignment, depth estimation, and adaptation to new surgical tasks with limited annotation. Overall, our results demonstrate that

thoughtful integration of accessible data streams and multimodal deep learning can deliver expert-level skill assessment, providing both high accuracy and practical flexibility across diverse real-world surgical settings.

Methods

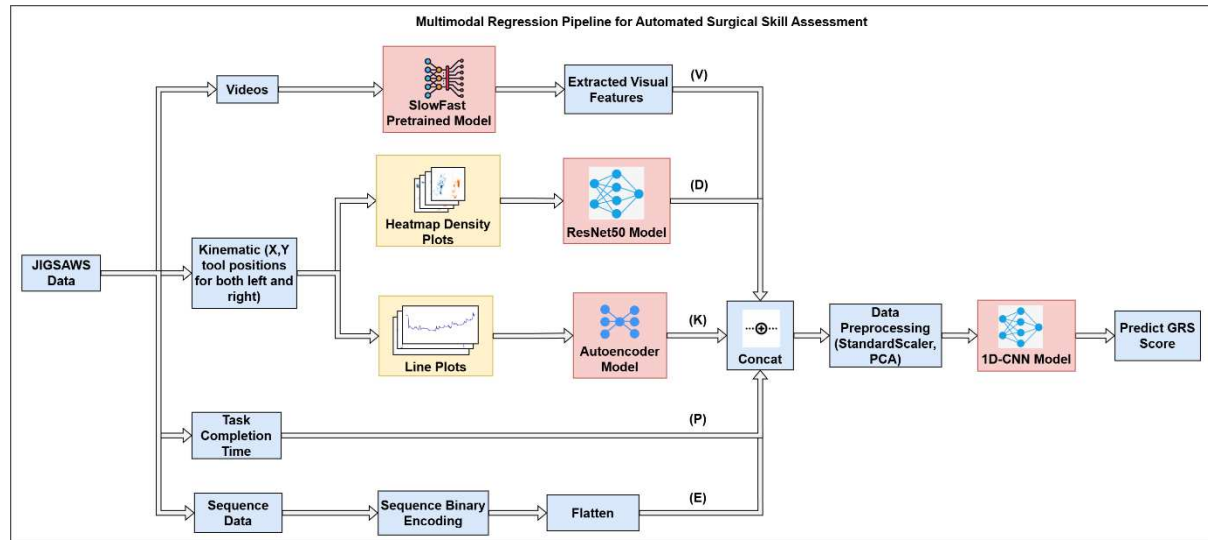


Figure 1. Schematic of the complete JIGSAWS feature extraction, fusion, and prediction pipeline.

Video features are extracted using a pretrained SlowFast model and represented as visual features (V). Tool kinematic trajectories are converted into line plots and encoded with an autoencoder to generate kinematic features (K), while heatmap density plots are processed with ResNet50 to generate semantic density features (D). Task completion time (P) and encoded gesture sequence information (E) are also included. The fused multimodal representation is standardized, reduced using principal component analysis, and input to a 1-dimensional convolutional neural network to predict the Global Rating Scale score.

A detailed schematic of the complete JIGSAWS feature extraction, fusion, and prediction pipeline is shown in Figure 1. This architecture demonstrates our multimodal design, where kinematic trajectories, heatmap density plots, task completion time, gesture event sequences, and video-based features are extracted and encoded via dedicated models (e.g., ResNet50, autoencoder, and SlowFast). The resulting features are concatenated, standardized, and passed to a 1D-Convolutional neural network (CNN) regression model to predict expert-assigned skill scores, enabling robust multimodal surgical skill assessment.

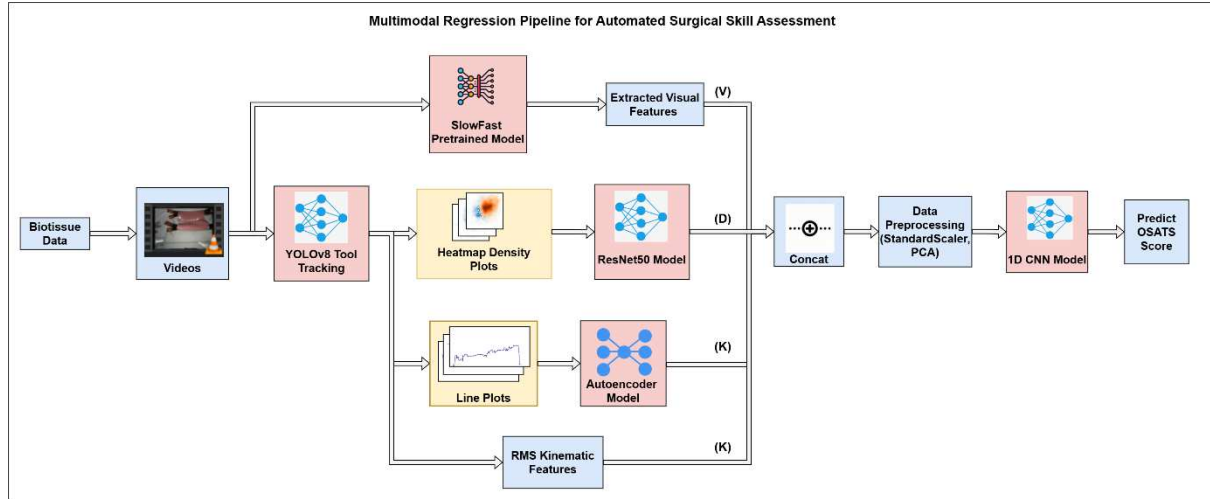


Figure 2. Schematic of the complete Biotissue dataset (GJ and PJ) feature extraction, fusion, and prediction pipeline

Surgical videos are processed using YOLOv8-based tool tracking to derive motion trajectories. Visual features (V) are extracted using a pretrained SlowFast model, semantic density features (D) are obtained from heatmap density plots using ResNet50, and kinematic features (K) are derived from line plots encoded by an autoencoder together with root mean square kinematic features. These features are concatenated, standardized, reduced by principal component analysis, and used by a 1-dimensional convolutional neural network to predict the Objective Structured Assessment of Technical Skills score.

Similarly, Figure 2 presents the architecture of our Biotissue pipeline, which adapts the same multimodal fusion strategy to robotic surgical training video. Here, video-derived tool tracking (YOLOv8), root mean square (RMS) kinematic features, heatmap density plots, and deep visual encodings (ResNet50 and SlowFast) are extracted from robotic surgical training videos. These representations are likewise combined and fed as an input to a 1D-CNN to predict OSATS scores, demonstrating the framework’s flexibility for both robotic and non-robotic surgical data.

The proposed method should be understood as a modular multimodal assessment framework rather than as a single identical architecture instantiated across datasets. The overall design principle is consistent across both settings: modality-specific feature extraction, fusion of complementary representations, and downstream skill prediction. However, the exact modality branches are adapted to the signals available in each dataset. In the Biotissue dataset, only surgical videos and OSATS scores were available, whereas direct robotic kinematic streams, event annotations, and task-completion metadata were not available for direct integration (or were not perfectly aligned for use in the present assessment pipeline). Moreover,

the videos contain practical imperfections, including pauses, variable zoom, and camera in/out motion. To address this, the framework derives motion-based tool-trajectory representations directly from video through segmentation, tracking, and trajectory extraction, thereby combining visual and video-derived kinematic information. In JIGSAWS, where richer modalities are directly available, these additional inputs can be incorporated explicitly within the same overall framework.

Tool Segmentation and Tracking

Instrument segmentation is performed using a YOLOv8²⁴ model, initialized with COCO²⁵ pretrained weights and fine-tuned on a curated set of manual and semiautomatically (SAM2-assisted²⁶) annotated frames.

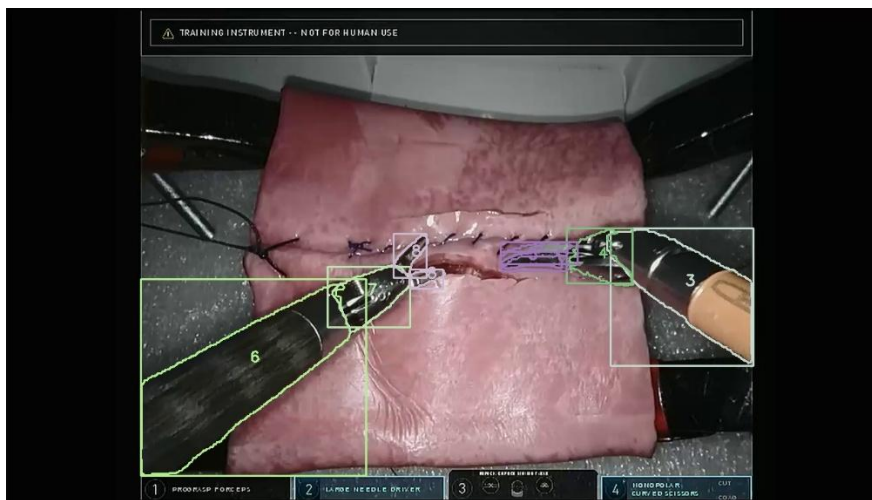


Figure 3. Segmentation output using SAM2-assisted annotation

Representative segmentation output generated using the SAM2-assisted annotation workflow on a surgical training frame from the Biotissue dataset. Colored outlines and masks indicate the segmented tool components, including wrist, shaft, and jaw regions, which were used to improve annotation quality for downstream YOLOv8 segmentation model training.

The introduction of SAM2 substantially streamlined our annotation workflow. By using YOLOv8’s bounding-box predictions as prompts, SAM2 generates precise, instance-level masks with minimal human oversight. Enhanced pixel-level accuracy and fewer spurious segmentations have reduced the need for extensive manual frame selection, allowing us to annotate many more frames in less time and rapidly scale our training set²⁷. Because SAM2 produces more precise and accurate tool masks (Figure 3), the labels used for training are of higher quality. This enables the YOLOv8 model to learn from clearer, more correct examples,

which improves its ability to identify and segment surgical tools in new images. The improvement is reflected in higher mAP@.50-.95 scores, a measure of how well the model detects and segments the tools. With these higher-quality labels and better scores, our final segmentation models become more robust and reliable for surgical video analysis.

For each video frame t , the model predicts bounding boxes and segmentation masks for all detected tool instances of the needle driver tool. The network is optimized to maximize both detection and segmentation performance, enabling accurate localization and pixel-wise delineation of multiple instruments per frame.

Detected tools are temporally linked across frames using BYTE Tracker, which associates bounding boxes based on spatial overlap and motion prediction. BYTE Tracker was selected for its demonstrated robustness to occlusion and real-time performance in surgical and cluttered scenes, as well as its superior tracking stability compared to prior approaches such as BotSort. These attributes make BYTE Tracker well-suited for the high-speed, multi-object tracking demands of surgical video analysis. For each tracked tool instance, the centroid $\mathbf{c}_{t,k} = (x_{t,k}, y_{t,k})$ of the corresponding mask is computed at each frame t . These centroid trajectories are then smoothed with a low-pass Butterworth filter (order n , cutoff frequency f_c), yielding noise-reduced, temporally coherent motion paths suitable for downstream kinematic analysis. After segmentation and tracking, the trajectory data were post-processed to reduce the impact of missing detections, brief occlusions, identity inconsistencies, and short track fragmentation. Most missing detections occurred during temporary jaw occlusion, while more stable subparts such as the wrist or shaft remained detectable. Short gaps also arose when instruments moved partially out of the field of view, for example during needle pickup or wider excursions associated with knot tying. Left-right inconsistencies were corrected using the shaft as the primary anchor, together with spatial proximity, bounding-box overlap, and short-term temporal consistency across adjacent frames. Although these discrepancies can introduce local noise into the trajectories, their downstream impact is reduced because the final kinematic representation is derived from processed motion patterns rather than isolated single-frame detections.

For the JIGSAWS dataset, we utilize the synchronized two-dimensional (x, y) Cartesian coordinates of the tool tips, sampled at 30 Hz from the three patient side manipulators (PSM1, PSM2, and PSM3), which are the robotic arms of the da Vinci surgical system responsible for instrument manipulation at the patient bedside⁴. In the standard JIGSAWS protocols, only PSM1 and PSM2 (corresponding to the left and right primary instrument arms) are used for most tasks, while PSM3 is available but typically not analyzed. All coordinates are normalized

to the $[0, 1]$ range to ensure model compatibility and to emphasize relative motion dynamics over absolute spatial positioning.

Kinematic Feature Encoding Using an Autoencoder

Each tool-tip trajectory $\{(x_t, y_t)\}_{t=1}^T$ is transformed into a pair of 1D line plot images: one representing the x-coordinate over time and the other the y-coordinate over time. These trajectory plot images, denoted as I_{traj} , are standardized and used as input to an autoencoder. This trajectory-to-image encoding strategy follows prior works, where x and y tool trajectories were similarly converted into plot-based representations and encoded with an autoencoder for surgical skill assessment^{11,27}. In the present study, the same formulation was retained to provide a compact and consistent motion representation for downstream multimodal fusion across datasets. The purpose of this autoencoder module is to learn a compact, robust, and noise-resilient representation of tool-tip motion dynamics, capturing the most salient kinematic patterns for downstream skill assessment. The autoencoder f_{enc} maps each plot image to a compact 256-dimensional latent feature vector $\mathbf{z}$, while the decoder f_{dec} reconstructs the original plot:

$$\mathbf{z} = f_{enc}(I_{traj}), \quad I'_{traj} = f_{dec}(\mathbf{z}) \quad (1)$$

The autoencoder Figure 4 is trained with a hybrid loss, combining mean squared error (MSE) and binary cross-entropy (BCE) to promote both pixel-level accuracy and structural fidelity. The resulting latent vectors provide robust, noise-resilient encodings of tool motion dynamics, which are incorporated as features for downstream skill assessment.

$$\mathcal{L}_{hybrid} = \text{MSE}(I_{traj}, I'_{traj}) + \lambda \cdot \text{BCE}(I_{traj}, I'_{traj}) \quad (2)$$

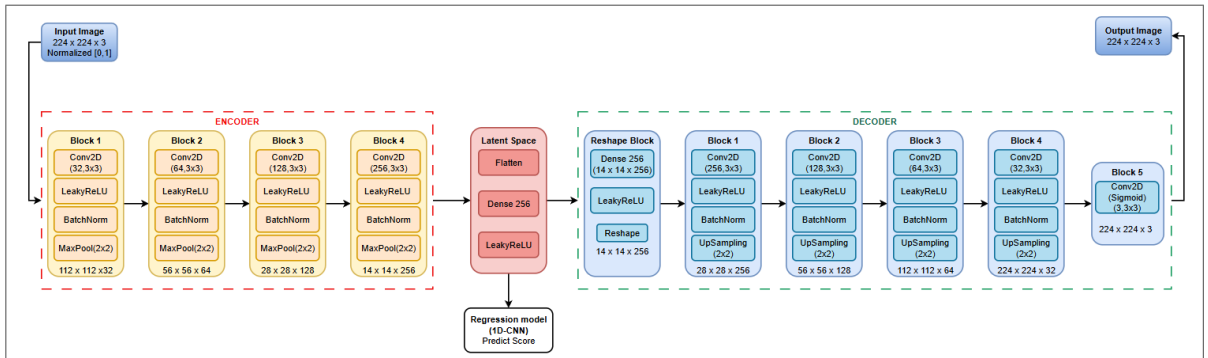


Figure 4. Autoencoder architecture

The autoencoder is used for kinematic feature encoding. The encoder consists of four convolutional blocks with batch normalization, LeakyReLU activation, and max pooling,

followed by a latent representation layer. The decoder reshapes the latent vector and reconstructs the input trajectory image through convolutional and upsampling blocks, ending with a sigmoid output layer. The latent representation is used as a compact feature encoding for downstream skill prediction.

Sequence Encoder Baselines

For the sequence encoder comparison, we additionally evaluated BiLSTM, TCN, and Transformer-based sequence encoders using the same video-derived tool trajectory inputs. Each input sequence had a fixed length of 512 timesteps with 12 trajectory channels corresponding to left and right tool-component coordinates. Each sequence encoder was trained within an autoencoder framework to reconstruct the input trajectory sequence using mean squared error loss. The encoder mapped each input trajectory sequence to a 128-dimensional latent representation, which was extracted and used as the kinematic trajectory feature in the downstream prediction pipeline. The models were trained using the AdamW optimizer, with early stopping based on validation reconstruction loss. The video feature branch, handcrafted kinematic statistics, downstream prediction model, cross-validation protocol, and evaluation metrics were kept unchanged across comparisons.

Spatio-Temporal Feature Extraction Using SlowFast-R50

To extract rich spatio-temporal representations from surgical video, we utilize the SlowFast-R50 neural architecture²⁸, pretrained on the Kinetics-400 dataset. This model employs a dual-pathway design: the *slow* pathway captures spatial context by analyzing sparsely sampled frames, while the *fast* pathway focuses on densely sampled, rapid temporal dynamics. This enables simultaneous modeling of detailed spatial features and subtle motion, making it particularly effective for complex activities such as surgery.

We considered alternative video backbones, including ViViT²⁹, during model development. However, because our pipeline required repeated embedding extraction from long surgical videos across multiple experiments, transformer-based video models were substantially more computationally demanding under the available hardware constraints. We therefore selected SlowFast-R50 as a practical compromise between computational efficiency and spatio-temporal representation quality. Its dual-pathway design was also well suited to surgical video, as it captures both slower contextual information and faster instrument-motion dynamics. We did not perform a systematic backbone ablation across all modalities in the present study, and

this is a design choice based on feasibility and task suitability rather than a definitive superiority claim.

Given a sequence of raw video frames $x_1, x_2, \dots, x_L$, the SlowFast model processes the input in parallel as follows:

$$\text{Slow pathway: } S_t = \text{Slow}(x_{1+\alpha \cdot (t-1)}), \quad t = 1, 2, \dots, T \quad (3)$$

$$\text{Fast pathway: } F_l = \text{Fast}(x_l), \quad l = 1, 2, \dots, L \quad (4)$$

where L is the total number of frames, $T = \lfloor \frac{L-1}{\alpha} \rfloor + 1$ is the number of slow pathway samples, and α is the temporal sampling ratio (e.g., $\alpha = 4$). Feature maps from both pathways are fused at multiple stages, enabling the network to combine global context and detailed motion.

For downstream modeling, we divide each video into fixed-length segments C_m , then preprocess (resize and normalize) and extract a high-dimensional spatio-temporal embedding for each segment:

$$\mathbf{f}_m = \text{SlowFast}(C_m) \quad (5)$$

where $\mathbf{f}_m \in \mathbb{R}^{2048}$ represents the embedding for segment m . Segment embeddings are concatenated in temporal order and zero-padded to standardize input size, preserving the full sequence of spatio-temporal features:

$$\mathbf{F}_{\text{video}} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_N] \quad (6)$$

where N is the number of segments per video.

We selected SlowFast over alternative video architectures such as X3D³⁰ and Video Vision Transformers (ViViT)²⁹ because SlowFast provides a strong balance of spatial context, motion detail, and computational efficiency. X3D is highly efficient but its uniform temporal sampling can miss subtle surgical motions, while ViViT requires substantially larger datasets and more computing to avoid overfitting. In contrast, SlowFast's dual-pathway design offers rich, fine-grained spatio-temporal representations, making it particularly effective for complex activity recognition in real-world scenarios.

Kinematic Feature Extraction – RMS Metrics and Path Length (Biotissue Dataset Only)

After smoothing Biotissue Gastrojejunostomy (GJ) tool-tip trajectories with a 24 Hz low-pass Butterworth filter, we quantified tool motion using RMS metrics and path length. Since precise tool-tip position data (for wrist, shaft, and jaw components) are available only in the Biotissue

dataset and not in JIGSAWS, these motion features were extracted exclusively for Biotissue. For each tool component, we computed:

$$\text{RMS}_{\text{velocity}} = \sqrt{\frac{1}{N} \sum_{i=1}^N v_i^2} \quad (7)$$

$$\text{RMS}_{\text{acceleration}} = \sqrt{\frac{1}{N} \sum_{i=1}^N a_i^2} \quad (8)$$

$$\text{RMS}_{\text{jerk}} = \sqrt{\frac{1}{N} \sum_{i=1}^N j_i^2} \quad (9)$$

Here, v_i , a_i , and j_i denote instantaneous velocity, acceleration, and jerk values, respectively, and N represents the total number of sampled data points.

Additionally, we computed the total path length traversed by each tool component across the entire trajectory, reflecting overall movement efficiency and precision:

$$\text{Path Length} = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (10)$$

All these features summarize motion smoothness, steadiness, and efficiency, key indicators of surgical skill. Calculations were limited to the Biotissue dataset, as the normalized coordinates of JIGSAWS precluded meaningful computation of these metrics. All extracted values were integrated into our multimodal feature set for skill prediction.

Semantic Feature Extraction – Heatmap Density Plots

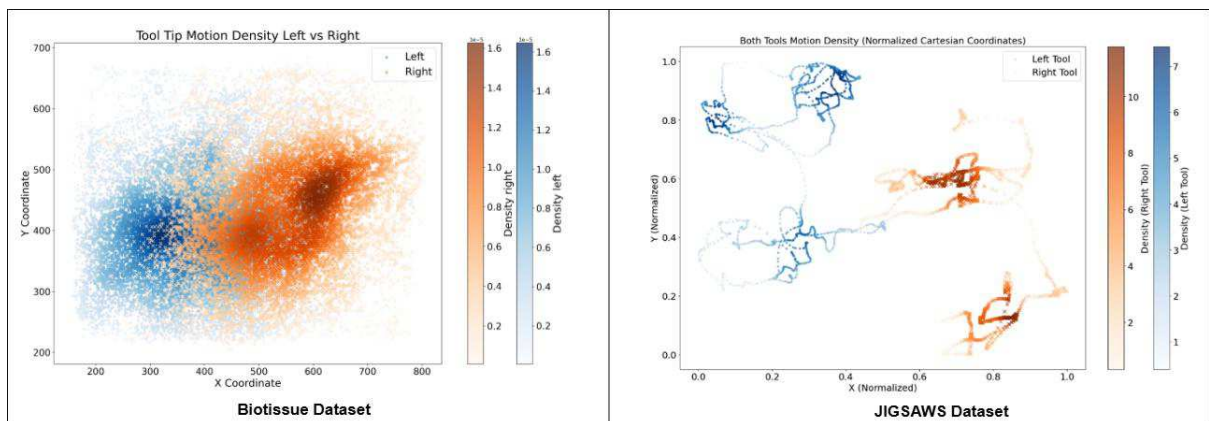


Figure 5. Heatmap density plots

Representative heatmap density plots generated from tool-tip motion trajectories in the Biotissue and JIGSAWS datasets. The left panel shows density maps for the Biotissue dataset in image-coordinate space, and the right panel shows density maps for the JIGSAWS dataset using normalized Cartesian coordinates. Blue corresponds to the left tool and orange corresponds to the right tool. These density maps are used to capture spatial motion distribution for semantic feature extraction.

To capture spatial behavioral patterns and tool motion distribution in surgical procedures, we generated two-dimensional heatmap density plots Figure 5 from tool-tip positions and extracted high-level semantic features using a pretrained ResNet-50 model. While the overall goal of semantic representation remained consistent across datasets, the methodology was tailored to data characteristics in the Biotissue and JIGSAWS datasets. In Biotissue, the presence of camera movement and frequent zoom changes can distort heatmap interpretation, making depth estimation a crucial direction for future work to accurately localize tool activity. In contrast, the JIGSAWS dataset benefits from a fixed camera, but the provided tool-tip positions are normalized Cartesian coordinates rather than absolute workspace locations, which may introduce additional noise. Despite these limitations, semantic features from heatmap representations provide useful contextual information in both datasets, offering notable value for the Biotissue dataset and contributing additional signal for JIGSAWS.

Event Sequence Encoding

When annotated surgical event or gesture sequences are available, we encode the gesture sequence for each video as a series of one-hot vectors. Let C denote the total number of gesture types, and N the number of gestures in a given video. Each gesture is mapped to a one-hot encoded vector of length C , indicating its category. The resulting $N \times C$ matrix is then flattened to yield a feature vector of length $N \times C$, which is concatenated with other video- and kinematic-derived features for downstream modeling. To create fixed-length feature vectors suitable for modeling, each array is zero-padded to match the maximum number of gesture events observed for that specific task. This encoding preserves the order of gestures within the video, providing a compact representation of the procedural sequence.

Encoding event or gesture sequences in this way is crucial because the type, order, and transitions of gestures can reflect both the efficiency and strategy of surgical performance. Procedural flow, repetition, and deviations from standard gesture sequences are often key indicators of technical skill, efficiency, and error. By capturing this sequential information, the

model can better distinguish skill levels, identify atypical performance patterns, and provide a richer context for automated skill assessment.

Example: One-hot encoding for gestures G12, G13, G14, G15, G11 ($C = 15, N = 5$)

Gesture order: G12, G13, G14, G15, G11

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
$C = 15$ (gesture types)															

$N = 5$ (number of gestures in this example; one row per gesture)

Flattened array: ($N \times C = 5 \times 15 = 75$ elements, row-wise concatenation)

[0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0]

Multimodal Feature Fusion and Skill Score Prediction

All feature vectors, including autoencoder latent codes, RMS/path metrics, semantic heatmaps, spatio-temporal SlowFast embeddings, and event sequence encodings, are concatenated to form a comprehensive multimodal representation:

$$\mathbf{X}_{\text{fused}} = (\mathbf{z}, \text{RMS}_{\text{velocity}}, \text{RMS}_{\text{acceleration}}, \text{RMS}_{\text{jerk}}, \text{Path length}, s, f_1, \dots, f_M, \mathbf{E}) \quad (11)$$

These features are standardized via z-score normalization, and dimensionality reduction (e.g., principal component analysis [PCA] retaining 99% variance) is applied prior to prediction.

PCA was used to compress the high-dimensional fused feature vectors into a tractable representation for downstream prediction. The subsequent 1D-CNN was used as a nonlinear prediction head over the reduced feature vector, and the PCA components were not interpreted as literal temporal steps. This design is also consistent with prior work using PCA-reduced representations followed by CNN-based modeling in other domains³¹, as well as with our own prior work using compact structured representations for CNN-based downstream prediction²⁷.

A 1D-CNN is trained to regress the expert-assigned skill score: $\hat{y} = R(\mathbf{X}_{\text{fused}})$

The model is optimized by minimizing the mean squared error loss:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12)$$

Evaluation Protocol

The performance of our model is evaluated using both the SCC (ρ) and mean absolute error (MAE). Specifically,

$$\rho = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \quad (13)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (14)$$

where d_i is the difference between the ranks of the true and predicted scores for sample i , and n is the number of samples.

For the JIGSAWS dataset, evaluation is performed using both 4-fold cross-validation and leave-one-user-out (LOUO) splits. For the Biotissue dataset, we employ 5-fold cross-validation on the GJ subset. For the expanded PJ subset, a separate stratified 5-fold cross-validation analysis was performed as a task-specific evaluation. In line with the recommendations discussed by Anastasiou et al.¹³, we emphasize that while the SCC is widely adopted as the primary evaluation metric, it only measures the agreement in ranking between predicted and true scores and does not reflect the actual prediction error. As highlighted by Contra-Sformer¹³, a model can achieve a high SCC despite large absolute prediction errors, especially on small test sets. To address this limitation, we report both MAE per task and the standard deviation of MAE across all folds, providing a more complete picture of model performance.

JIGSAWS Dataset

For benchmarking, we used the public JIGSAWS dataset⁴, comprising 206 stereo endoscopic videos of three simulated robotic tasks: suturing (78 videos), needle-passing (56), and knot-tying (72). Recordings were captured at 30 Hz and 640×480 resolution from two viewpoints; we analyzed only the “capture1” (left) videos, yielding 39 suturing, 28 needle-passing, and 36 knot-tying clips.

Each trial includes an expert-assigned GRS (range: 6–30). JIGSAWS provides frame-level gesture annotations (G1–G15; see Table 12) and synchronized kinematic data, enabling

detailed motion analysis and robust model supervision. Gesture sequences were one-hot encoded for input features. Unlike the Biotissue dataset, JIGSAWS enables fine-grained procedural segmentation, though the absence of stereo camera calibration limits 3D reconstruction accuracy.

Table 12: JIGSAWS gesture index and description.

Gesture index	Gesture description
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to center with needle in grip
G6	Pulling suture with left hand
G7	Re-grip needle
G8	Orienting needle
G9	Using the right hand to help tighten suture
G10	Loosening more suture
G11	Dropping suture at end and moving to end points
G12	Reaching for needle with left hand
G13	Making C loop around right hand
G14	Reaching for suture with right hand
G15	Pulling suture with both hands

For JIGSAWS, 4-fold cross-validation was performed (as no predefined folds exist), using balanced splits by task. To ensure a fair comparison, we followed the commonly adopted splits described in Tang et al.³² and Liu et al.¹⁰. The evaluation is based on SCCs between model predictions and expert ratings.

Biotissue Dataset

Our primary clinical dataset comprises 205 high-resolution videos (1280×720 pixels) of anastomoses performed on synthetic hydrogel Biotissue models (Lifelike Biotissue, Inc.) using the DaVinci Xi robot, subdivided into three categories: 33 pancreaticojejunostomy, 44

hepaticojejunostomy, and 128 gastrojejunostomy recordings. All procedures were performed by surgical trainees under standard institutional protocols at UT Southwestern Medical Center. The video data were collected as part of a quality improvement project reviewed by the Institutional Review Board at UT Southwestern Medical Center (IRB No. Y1-21-507), which determined that the project did not meet the definition of human subjects research. For downstream skill assessment, the GJ analysis used the same 128 videos, whereas the PJ analysis used an expanded cohort of 170 videos, including the original 33 PJ recordings and additional videos collected subsequently.

Simulation-based surgical training has become an essential component of many curricula, enabling trainees to practice in a risk-free controlled environment³³⁻³⁵. At UT Southwestern, we have developed and validated a proficiency-based robotic surgery simulation curriculum that combines a virtual reality platform (SimNow) with dry-lab exercises, followed by procedure-specific Biotissue drills³⁶⁻⁴⁰. These performances are recorded and subsequently assessed using global (OSATS and the Global Evaluative Assessment of Robotic Skills) measures as well as task-specific metrics and gestures^{1,41-43}. While effective, these manual evaluations are resource-intensive and time-consuming, which has limited their scalability. One of our prior video-based assessment studies demonstrated that technical proficiency in simulated gastrojejunostomy drills is closely correlated with operative performance in real clinical cases. Importantly, surgical trainees who achieved higher performance scores in simulation experienced higher performance in the operating room and a lower incidence of clinically relevant delayed gastric emptying¹⁹. All Biotissue videos were evaluated using the OSATS scoring framework. Because GJ, PJ, and HJ represent distinct procedural tasks with different workflow, motion characteristics, technical demands, and sources of difficulty, we treated downstream skill assessment as a task-specific prediction problem rather than training a single unified predictor across heterogeneous procedures. In the present study, we report results for both the GJ and PJ subsets. HJ was not included in the downstream skill assessment experiments because the number of available HJ videos was too limited for a stable evaluation. By contrast, the segmentation model was developed using data from multiple procedure types to improve the robustness of tool localization across the dataset.

OSATS-based video assessment for the Biotissue dataset was performed independently by two research fellows who had completed a standardized institutional training framework for performance assessment within our validated robotic surgery curriculum. Both graders applied the OSATS rubric independently. Inter-rater reliability was quantified using the intraclass correlation coefficient (ICC) and reported in our prior works^{19,20}, which demonstrated good

agreement (ICC = 0.86, 95% CI: 0.69–0.93; $p < 0.001$) for GJ and ICC of 0.74 (0.6–0.84; $p < 0.05$) for the PJ dataset. The two graders were blinded to each other's evaluations, and the final score assigned to each video was computed as the average of the two ratings.

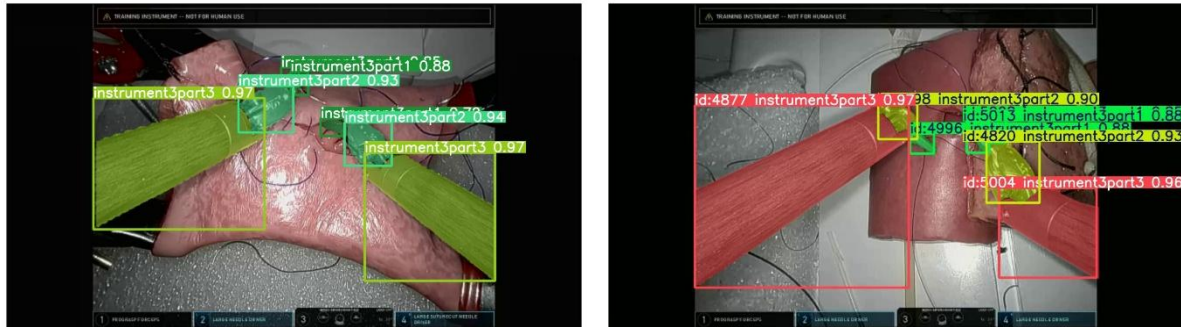


Figure 6. Representative YOLOv8-based tool tracking outputs in the Biotissue dataset. Representative YOLOv8 tool tracking outputs from the Biotissue dataset. The left panel shows gastrojejunostomy examples and the right panel shows pancreatojejunostomy examples. The figure illustrates tracked needle driver components in surgical video frames and highlights the visual outputs used for downstream trajectory extraction and motion analysis.

A total of 1239 frames spanning all procedure types were manually annotated by trained medical interns and surgical staff, with each needle driver segmented into wrist, shaft, and jaw (classified as left or right; see Figure 6). To augment the training set, 152 pseudo-labeled images were generated using a semi-automated workflow: initial YOLOv8²⁴ segmentation followed by SAM2²⁶ refinement. Further augmentation was limited due to diminishing accuracy gains. The final dataset comprised 1391 training, 153 validation, and 308 test images. YOLOv8 was initialized with COCO-pretrained weights²⁵ and fine-tuned on the Biotissue dataset. Segmentation performance was evaluated using mean average precision (mAP@0.50–0.95), which averages precision across IoU thresholds and is the standard metric for segmentation quality in computer vision.

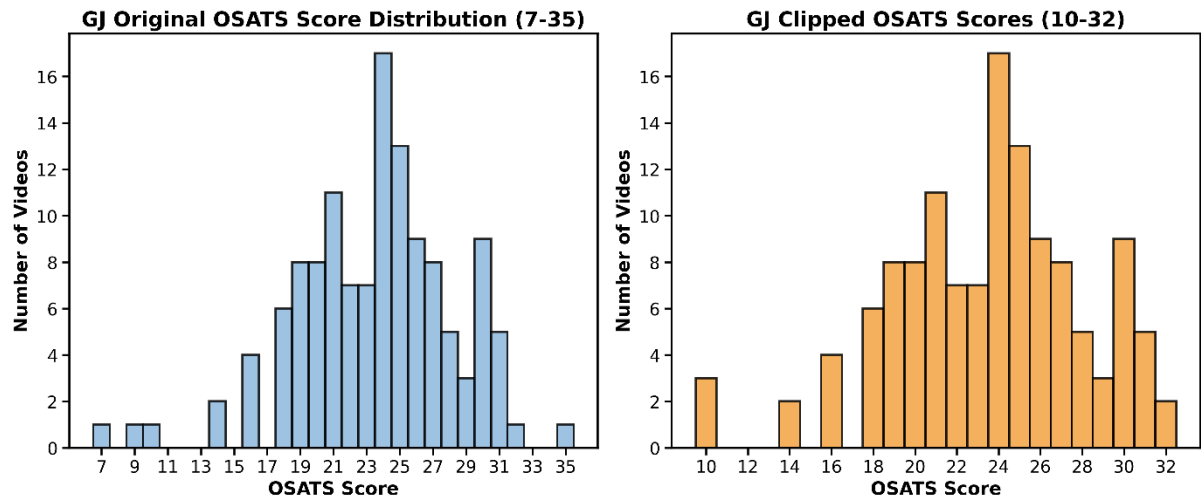


Figure 7. Score distribution for gastrojejunostomy subset. Distribution of Objective Structured Assessment of Technical Skills scores for the gastrojejunostomy subset of the Biotissue dataset. The left panel shows the original score distribution, and the right panel shows the clipped score distribution used for downstream analysis after exclusion of extreme values. The horizontal axis indicates score, and the vertical axis indicates number of videos.

We focused skill assessment on the gastrojejunostomy procedure subset, which includes 128 recordings with OSATS scores (range: 7–35) assigned by board-certified surgeons across seven performance domains. To mitigate outlier effects, scores were clipped to 10–32, excluding the lowest and highest 10% of samples (see Figure 7 for score distributions). We also evaluated skill assessment on an expanded pancreaticojejunostomy (PJ) subset comprising 170 videos with OSATS scores (range: 7–35) (see Figure 8 for score distributions).

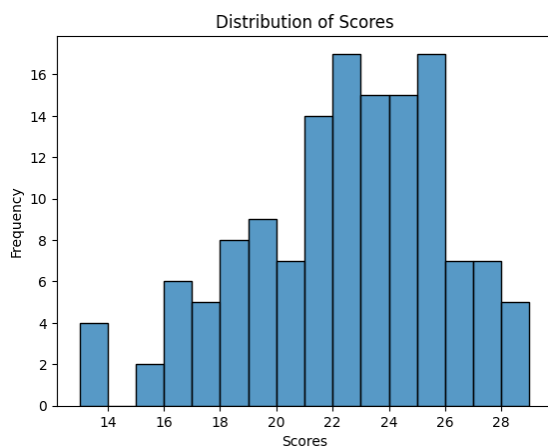


Figure 8. Score distribution for pancreaticojejunostomy subset. Distribution of Objective Structured Assessment of Technical Skills scores for the pancreaticojejunostomy subset of the Biotissue dataset. The histogram shows the score frequency

across videos included in the pancreateojejunostomy analysis. The horizontal axis indicates score and the vertical axis indicates frequency.

For the Biotissue experiments, model performance was evaluated separately for each task-specific subset. The GJ subset was evaluated using 5-fold cross-validation, whereas the PJ subset was evaluated using stratified 5-fold cross-validation to better preserve the distribution of OSATS scores across folds and reduce fold-level imbalance. Performance was assessed using SCC between predicted and expert-assigned scores.

Preprocessing

All multimodal features and target scores were systematically preprocessed to ensure normalized input scales, reduced dimensionality, and stable model convergence. For each dataset, the complete feature set included autoencoded kinematic features, RMS motion statistics (Biotissue only), semantic heatmap features, spatio-temporal SlowFast-R50 embeddings, and one-hot encoded gesture events (JIGSAWS only). All feature vectors were standardized using zero-mean, unit-variance scaling to prevent any single feature type from dominating the learning process.

Target variables (GRS for JIGSAWS and OSATS for Biotissue) were independently scaled to the $[0, 1]$ range using Min-Max normalization.

In JIGSAWS, gesture event sequences were converted to one-hot encoded matrices, flattened, and concatenated with other features prior to scaling and PCA. For videos of varying lengths, zero-padding was employed to achieve consistent input dimensions for batch processing in the 1D-CNN. All fitted preprocessing scalers and PCA models were saved for each cross-validation fold to ensure reproducibility and robust evaluation.

To further reduce redundancy and computational overhead, PCA was applied to the standardized features, retaining 99% of the total variance. Table 13 provides a summary of the feature dimensions before and after this reduction for both datasets.

Table 13: Summary of multimodal feature dimensions before and after principal component analysis (PCA). Abbreviations: D, semantic density features; E, event sequence; K, kinematic features; P, task completion time; V, video features.

Dataset configuration	Initial features	Post-PCA (n -comp)	Variance
Biotissue GJ (V, K, D)	119,170	~92	99%
Biotissue PJ (V, K)	101,045	~119	99%
JIGSAWS (V, K, E, P, D)	7,077	~29	99%

1D-CNN Training

The fused multimodal features, including spatio-temporal, semantic, and kinematic modalities, were used to train a 1D-CNN to predict surgical skill scores (GRS for JIGSAWS, OSATS for Biotissue). For model input, the PCA-transformed features representing 92 principal components for Biotissue GJ, 119 for Biotissue PJ and 29 for JIGSAWS were used as compact reduced representations for the Conv1D layers. This allows the network to identify hierarchical correlations between components that capture the most significant variance across the fused modalities (V, K, E, D, P). Figure 9 illustrates the network architecture.

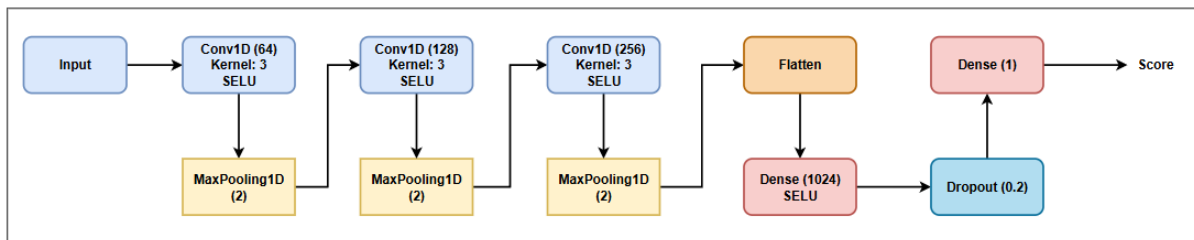


Figure 9. 1D-CNN architecture diagram.

Architecture of the 1-dimensional convolutional neural network used for skill score prediction. The model consists of three convolutional layers with kernel size 3 and SELU activation, each followed by max pooling, then flattening, a dense layer with 1024 units, dropout, and a final dense output layer for regression.

The 1D-CNN model architecture, as summarized in Table 14, consisted of:

- Three Conv1D layers (with 64, 128, and 256 filters; kernel size 3; SELU activation; and L2 regularization)
- MaxPooling1D layers after each convolution
- A flattened layer
- A fully connected dense layer (1024 units, SELU activation) followed by dropout (rate 0.2)
- An output dense layer (1 unit) for regression

Input features were reshaped (to samples, components, and channels) for the Conv1D layers, with PCA components treated as timesteps and the channel dimension set to one. Model training employed the Adam optimizer for GJ and JIGSAWS, and AdamW for PJ (learning rate = 0.0002) using mean squared error loss, for up to 1000 epochs, with early stopping (patience = 200) based on validation loss. Five-fold cross-validation was used for the Biotissue dataset and 4-fold cross-validation for JIGSAWS, with all trained models, scalers, and PCA objects saved per fold. Validation predictions were aggregated across folds to assess the model’s reliability in predicting expert-assigned surgical skill scores.

Table 14: Summary of model architectures and main parameters. Abbreviations: 1D-CNN, 1-dimensional convolutional neural network; IoU thr, Intersection over union threshold; LR, Learning Rate; MSE, mean squared error.

Model	Parameter	Value
YOLOv8-Seg Medium	Epochs	1000
	Batch size	32
	Image size	640×640
	Optimizer	SGD
	Initial LR	0.001
	Final LR	0.01
	Momentum	0.937
	Weight decay	0.0005
	Patience	400
	IoU thr.	0.7
Autoencoder	Type	Conv AE
	Input shape	224×224×3
	Encoder filters	32, 64, 128, 256
	Dropout	0.2
	Latent dim	256
	Output act.	Sigmoid
	Optimizer	Adam
	LR	1×10^{-6}
	Loss	MSE

	Batch size	1
	Epochs	300
	Patience	30
1D-CNN	Activation	SELU
	Optimizer	Adam (GJ and Jigsaws), AdamW (PJ)
	LR	0.0002
	L2	1×10^{-5}
	Loss	MSE
	Metrics	MAE
	Epochs	1000
	Batch size	1
	Patience	200

All experiments were conducted on a high-performance workstation equipped with an AMD Ryzen Threadripper PRO 5955WX 16-core processor (4.0 GHz, 128 GB RAM) and two NVIDIA RTX 4090 GPUs (24 GB VRAM each).

Data availability

- **Biotissue Dataset:** The data supporting this study's findings are not publicly available due to privacy and confidentiality restrictions. However, the datasets can be obtained from the corresponding author upon reasonable request, subject to approval and compliance with relevant data protection regulations.
- **JIGSAWS Dataset:** Details regarding access and the request procedure are available at: <https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws/>

Code availability

The code for reproducing the experimental results using the released, pre-extracted feature sets is available at: <https://doi.org/10.5281/zenodo.18602718>

Acknowledgments

This study received no funding. The authors thank Dave Primm of the UT Southwestern Department of Surgery for help in editing this article.

Author contributions

- G.S., S.M.K and H.P.N conceived the idea.
- S.M.K. applied the methodology and carried out coding, experiments, data preprocessing, data analysis, and manuscript drafting.
- H.P.N. contributed to reviewing and editing the manuscript.
- S.G.P., S.C., A.A., and A.A.A. graded Biotissue surgical videos.
- K.G. and S.G.P. annotated the Biotissue dataset for segmentation model training.
- D.S., P.P., and H.Z. contributed to manuscript review and critical feedback.
- G.S. contributed to manuscript review, editing, and overall supervision.
- All authors reviewed and approved the final manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

References

- 1 Martin, J. *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *British journal of surgery* **84**, 273-278 (1997).
- 2 Doyle, J. D., Webber, E. M. & Sidhu, R. S. A universal global rating scale for the evaluation of technical skills in the operating room. *The American journal of surgery* **193**, 551-555 (2007).
- 3 Beard, J. D. Assessment of surgical skills of trainees in the UK. *The Annals of The Royal College of Surgeons of England* **90**, 282-285 (2008).
- 4 Gao, Y. *et al.* in *MICCAI workshop: M2cai*. 3.
- 5 Lea, C., Vidal, R., Reiter, A. & Hager, G. D. in *Computer Vision – ECCV 2016 Workshops*. (eds Gang Hua & Hervé Jégou) 47-54 (Springer International Publishing).
- 6 Twinanda, A. P. *et al.* Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**, 86-97 (2016).
- 7 Zia, A. *et al.* in *Medical Image Computing and Computer-Assisted Intervention--MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I* 18. 430-438 (Springer).
- 8 Zia, A. *et al.* Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International journal of computer assisted radiology and surgery* **11**, 1623-1636 (2016).
- 9 Funke, I., Mees, S. T., Weitz, J. & Speidel, S. Video-based surgical skill assessment using 3D convolutional neural networks. *International journal of computer assisted radiology and surgery* **14**, 1217-1225 (2019).
- 10 Liu, D. *et al.* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9522-9531.
- 11 Yanik, E., Kruger, U., Intes, X., Rahul, R. & De, S. Video-based formative and summative assessment of surgical tasks using deep learning. *Sci Rep* **13**, 1038 (2023). <https://doi.org/10.1038/s41598-022-26367-9>
- 12 Yu, X., Rao, Y., Zhao, W., Lu, J. & Zhou, J. in *Proceedings of the IEEE/CVF international conference on computer vision*. 7919-7928.
- 13 Anastasiou, D., Jin, Y., Stoyanov, D. & Mazomenos, E. Keep your eye on the best: contrastive regression transformer for skill assessment in robotic surgery. *IEEE Robotics and Automation Letters* **8**, 1755-1762 (2023).
- 14 Majeedi, A., Gajjala, V. R., GNVV, S. S. S. N. & Li, Y. in *European Conference on Computer Vision*. 143-161 (Springer).
- 15 Bai, Y. *et al.* in *European conference on computer vision*. 422-438 (Springer).
- 16 Li, Z., Gu, L., Wang, W., Nakamura, R. & Sato, Y. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 410-420 (Springer).
- 17 Zia, A. & Essa, I. Automated surgical skill assessment in RMIS training. *International Journal of Computer Assisted Radiology and Surgery* **13**, 731-739 (2018). <https://doi.org/10.1007/s11548-018-1735-5>
- 18 Youssef, S. C. *et al.* Learning surgical skills through video-based education: a systematic review. *Surgical Innovation* **30**, 220-238 (2023).
- 19 Abreu, A. A. *et al.* From the simulation lab to the operating room: simulation performance predicts intraoperative performance in robotic gastrojejunostomy. *Surg Endosc* **38**, 5967-5973 (2024). <https://doi.org/10.1007/s00464-024-11035-y>
- 20 Palacios, S. G. *et al.* The reliability of task-specific metrics to assess resident performance on a robotic simulated pancreatojejunostomy. *Global Surgical Education-Journal of the Association for Surgical Education* **4**, 76 (2025).
- 21 Wang, T., Wang, Y. & Li, M. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* 23. 668-678 (Springer).
- 22 Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735-1780 (1997).
- 23 Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* **18**, 602-610 (2005).
- 24 Jocher, G., Chaurasia, A. & Qiu, J. YOLO by Ultralytics (Version 8.0. 0)[Computer software]. *YOLO by Ultralytics (Version 8.0. 0)[Computer software]* (2023).

- 25 Lin, T.-Y. *et al.* in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*. 740–755 (Springer).
- 26 Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R. & Ryali, C. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024).
- 27 Khairnar, S. M. *et al.* Machine Learning-Based Automated Assessment of Intracorporeal Suturing in Laparoscopic Fundoplication. (2024). <https://doi.org/10.48550/arxiv.2412.16195>
- 28 Feichtenhofer, C., Fan, H., Malik, J. & He, K. SlowFast Networks for Video Recognition. *CoRR abs/1812.03982* (2018).
- 29 Arnab, A. *et al.* ViViT: A Video Vision Transformer. *CoRR abs/2103.15691* (2021).
- 30 Feichtenhofer, C. X3D: Expanding Architectures for Efficient Video Recognition. *CoRR abs/2004.04730* (2020).
- 31 Agarwalla, A., Dileep, D., Jyothsana, P., Unnikrishnan, P. & Thirumala, K. in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. 1–5 (IEEE).
- 32 Tang, Y. *et al.* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9839–9848.
- 33 Dulan, G. *et al.* Developing a comprehensive, proficiency-based training program for robotic surgery. *Surgery* **152**, 477–488 (2012).
- 34 Arain, N. A. *et al.* Comprehensive proficiency-based inanimate training for robotic surgery: reliability, feasibility, and educational benefit. *Surgical endoscopy* **26**, 2740–2745 (2012).
- 35 Stefanidis, D. & Heniford, B. T. The formula for a successful laparoscopic skills curriculum. *Archives of surgery* **144**, 77–82 (2009).
- 36 Farah, E. *et al.* Adapting to a robotic era: the transferability of open and laparoscopic skills to robotic surgery. *Journal of Surgical Education* **80**, 1868–1876 (2023).
- 37 Alterio, R. E. *et al.* Developing a robotic surgery curriculum: selection of virtual reality drills for content alignment. *Journal of Surgical Research* **283**, 726–732 (2023).
- 38 Radi, I. *et al.* Feasibility, effectiveness and transferability of a novel mastery-based virtual reality robotic training platform for general surgery residents. *Surgical endoscopy* **36**, 7279–7287 (2022).
- 39 Nagaraj, M. B. *et al.* Feasibility and skill acquisition for a novel proficiency-based robotic surgery curriculum: a randomized medical student pilot. *Global Surgical Education-Journal of the Association for Surgical Education* **1**, 67 (2022).
- 40 Hogg, M. E. *et al.* Mastery-based virtual reality robotic simulation curriculum: the first step toward operative robotic proficiency. *Journal of surgical education* **74**, 477–485 (2017).
- 41 Al Abbas, A. I. *et al.* The development of task-specific metrics for grading the robotic gastrojejunostomy in robotic pancreaticoduodenectomy. *Global Surgical Education-Journal of the Association for Surgical Education* **3**, 4 (2023).
- 42 Sánchez, R. *et al.* Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). *Journal of robotic surgery* **10**, 227–231 (2016).
- 43 Haque, T. F. *et al.* An assessment tool to provide targeted feedback to robotic surgical trainees: development and validation of the end-to-end assessment of suturing expertise (EASE). *Urology practice* **9**, 532–539 (2022).
- 44 Pan, J.-H., Gao, J. & Zheng, W.-S. in *Proceedings of the IEEE/CVF international conference on computer vision*. 6331–6340.
- 45 Gao, J. *et al.* in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. 222–238 (Springer).