KIDA: Kinematic-Intent Dual-path Alignment for Surgical Error Detection

Yuxuan Luo[2,†][0009−0008−4915−0065], Kaixiang Yang[1,2,†][0009−0008−9160−8498], Yuxi Li[1,2,†][0009−0003−7143−2986], Wei Fang[3][0009−0001−9576−4673], Qiang Li[1(✉)][0000−0002−9815−4432], and Zhiwei Wang[1(✉)][0000−0002−1612−8573]

[1] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

[2] College of Life Science and Technology, Huazhong University of Science and Technology

[3] Wuhan United Imaging Surgical Co., Ltd.

[†] Co-first authors, [(✉)] Corresponding authors

{yuxuanluo, kxyang, liyuxi9, liqiang8, zwwang}@hust.edu.cn

Abstract. Automated Surgical Error Detection is critical for intraoperative safety and objective skill assessment in robotic-assisted surgery. Although major errors are visually apparent, subtle execution-level errors reflect early control degradation and often resemble normal maneuvers, posing a significant challenge for reliable detection. Although existing methods explore various spatio-temporal and multi-modal features, they formulate error detection as implicit discrimination. Consequently, they struggle to resolve the visual ambiguity of subtle errors without an explicit reference of correct execution. In this paper, we propose **KIDA**, a **K**inematic-**I**ntent **D**ual-path **A**lignment framework that enhances the discrimination of subtle surgical errors by explicitly aligning actual execution with clinically grounded intents. KIDA leverages kinematic information to more comprehensively capture the current motion state, and incorporates predefined textual descriptions based on clinical standards to serve as an explicit reference for correct execution. Specifically, the Kinematic Execution Branch (KEB) processes lightweight kinematic motion dynamics to provide subtle execution details about the ongoing surgical action. The Clinically-Guided Intent Branch (CIB) utilizes semantic intent derived from clinical textual descriptions to provide explicit guidance, dynamically evaluating whether the motion-aware visual features conform to the intended standard maneuver. To further improve error discrimination, we introduce an Intent-Execution Alignment (IEA) optimization strategy, which explicitly enforces consistency for normal actions while actively decoupling subtle errors from the semantic intent to enlarge the feature margin. Experiments on the SAR-RARP50 dataset demonstrate that KIDA achieves state-of-the-art performance, improving AUC by 2.28% and AP by 4.86% while demonstrating generalizability on the JIGSAWS benchmark with a 0.84% AUC improvement. Code is available at https://github.com/yxluo8/KIDA.

Keywords: Surgical Error Detection · Robotic-Assisted Surgery · Kinematic-Intent Alignment.

1 Introduction

The rapid adoption of Robotic-Assisted Surgery (RAS) demands reliable automated Surgical Error Detection (SED) to ensure intraoperative patient safety and provide objective skill assessment [24,13]. While catastrophic mistakes such as severe hemorrhage or direct organ injury [2] may be visually apparent, many clinically relevant execution errors are considerably more subtle [10]. These errors often arise from improper motion control, repeated ineffective attempts, or unstable manipulations like inadequate suture tension. If left uncorrected, these seemingly minor deviations can accumulate and cascade into severe postoperative complications, such as anastomotic urinary leakage [4,15]. Furthermore, because these faults remain visually similar to valid but complex procedural variations, such ambiguity poses a fundamental challenge for reliable automated detection [20,7].

Early SED research primarily focused on structured dry-lab settings [1]. In these controlled environments, Li et al. [12] proposed runtime error detection using contrastive metric learning on trajectories, while Shao et al. [18] introduced the Chain-of-Gesture (COG) framework, employing gesture-level prompts to enhance temporal reasoning. With the advent of clinically realistic datasets such as SAR-RARP50 [16], research shifted toward untrimmed in-vivo procedures characterized by long durations and substantial inter-surgeon variability. To address the complex dynamics of extended surgical phases, recent methods rely heavily on aggregating temporal features over long windows to smooth out execution noise. For example, Sirajudeen et al. [19] introduced MS-TCN++ cascade for refined frame-level predictions, and Xu et al. [21] proposed SEDMamba, utilizing Selective State Space Models (SSMs) [8,9] to efficiently capture long-range visual dependencies.

Despite the progression from controlled dry-lab settings to complex in-vivo scenarios and the exploration of diverse multi-modal cues and long-range temporal modeling [12,18,19,21], reliably detecting subtle execution errors remains challenging. Most existing approaches enrich spatio-temporal representations and directly predict error probabilities from observed features, resulting in a predominantly appearance-driven formulation. However, subtle execution errors often share highly similar visual and motion characteristics with valid surgical maneuvers, leading to substantial overlap in feature space. When discrimination depends solely on feature-level differences, distinguishing genuine faults from acceptable procedural variations becomes inherently difficult.

To address this limitation, we reformulate SED from an implicit feature-level classification into an explicit measurement of intent-execution consistency. Based on this perspective, we propose **KIDA** (**K**inematic-**I**ntent **D**ual-path **A**lignment), a framework that determines surgical errors by comparing observed execution with clinically defined normative behavior. KIDA models this compar-

ison through two complementary branches. The Kinematic Execution Branch (KEB) first enhances the representation of ongoing surgical motion by leveraging lightweight kinematic cues, providing fine-grained information about actual execution beyond appearance-based features. Subsequently, the Clinically-Guided Intent Branch (CIB) introduces textual descriptions derived from clinical standards to dynamically query the visual representations, effectively evaluating whether the ongoing execution conforms to the intended maneuver. For furthur discrimination, we design an Intent-Execution Alignment (IEA) optimization. IEA explicitly regulates the fused representations by enforcing sparsity to maintain execution stability for normal maneuvers, while actively minimizing the semantic similarity between the unconstrained kinematic discrepancy and the correct intent, thereby pushing erroneous executions away from the normative feature space. Our main contributions are summarized as follows:

- We reformulate Surgical Error Detection as measuring execution consistency with clinically grounded intent, shifting from purely appearance-driven discrimination to intent-execution alignment.
- We propose KIDA, which integrates a Kinematic Execution Branch (KEB) and a Clinically-Guided Intent Branch (CIB) to model actual motion and normative intent, optimized via an auxiliary Intent-Execution Alignment (IEA) regularization to explicitly enforce strict consistency while actively decoupling subtle errors from the semantic intent.
- Extensive experiments on the highly challenging *in-vivo* SAR-RARP50 dataset demonstrate the superiority of our framework, achieving state-of-the-art (SOTA) results by improving AUC by 2.28% and AP by 4.86%. Furthermore, KIDA demonstrates generalizability on the JIGSAWS, yielding a 0.84% AUC improvement.

2 Method

Fig. 1 illustrates the overall architecture of the proposed KIDA framework. Unlike existing approaches that directly map visual features to implicit temporal spaces, KIDA structurally enforces explicit alignment between physical execution and clinical intent. The framework is built upon an spatio-temporal backbone and incorporates three core components: (i) the Kinematic Execution Branch (KEB); (ii) the Clinically-Guided Intent Branch (CIB); and (iii) the Intent-Execution Alignment (IEA) optimization. In the following sections, we describe each component in detail.

2.1 Backbone Architecture

We adopt the spatiotemporal modeling framework of SEDMamba [21], which follows a sequential pipeline of visual feature extraction followed by temporal modeling. Given an input video $V = \{x_t\}_{t=1}^{T}$, we first employ DINOv2-Giant with registers [14] as the visual encoder to extract frame-level representations

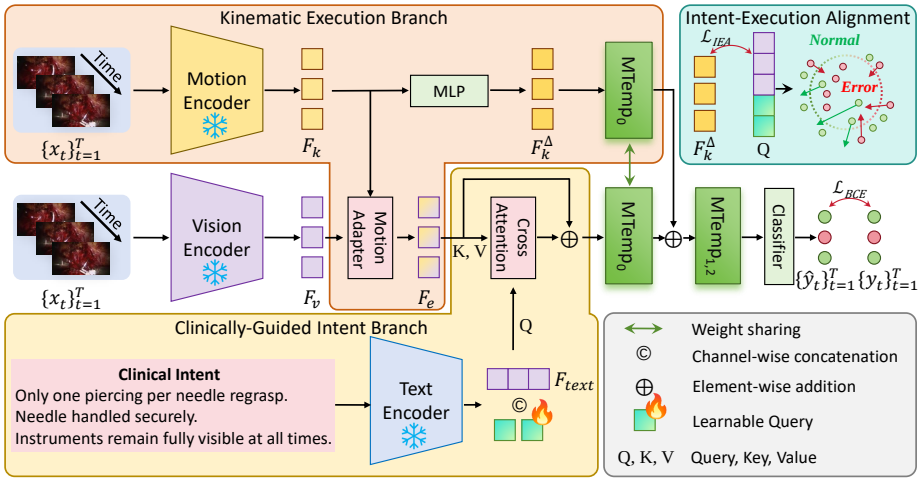

Fig. 1. Overview of the proposed KIDA framework. It comprises three components: (i) the Kinematic Execution Branch (KEB), which enhances visual features with kinematic cues; (ii) the Clinically-Guided Intent Branch (CIB), which constructs a textual reference of correct execution from clinically grounded textual descriptions; and (iii) the Intent-Execution Alignment (IEA) optimization, which enforces consistency between execution and intent for improved error discrimination.

$F_v \in \mathbb{R}^{T\times C}$, where $C = 1000$. The extracted features are then fed into a Mamba-based temporal module (MTemp) from [21] to capture long-range temporal dependencies and obtain spatiotemporal representations. Finally, a convolutional layer is applied to generate frame-level anomaly scores for surgical error detection.

2.2 Kinematic Execution Branch

While the backbone captures general spatiotemporal dynamics, it remains limited in distinguishing subtle execution errors that exhibit highly similar visual patterns [2]. We introduce the Kinematic Execution Branch (KEB), which explicitly injects motion cues to enhance fine-grained physical modeling.

Given visual features $F_v \in \mathbb{R}^{T\times C}$, we extract kinematic motion signals $F_k \in \mathbb{R}^{T\times 6}$ using OpenCV-based optical flow. KEB integrates motion information to obtain the enhanced execution feature F_e:

$$F_e = F_v + \phi(F_k) \odot \sigma(\psi(F_k)), \tag{1}$$

where $\phi(\cdot)$ projects low-dimensional kinematic signals into the visual embedding space, $\psi(\cdot)$ learns motion saliency for adaptive gating, $\odot$ denotes element-wise multiplication, and σ is the sigmoid activation.

To further preserve micro-level kinematic instabilities that may be smoothed during temporal aggregation, we additionally derive a residual from the raw

motion signals:

$$F_k^{\Delta} = \texttt{MLP}(F_k), \tag{2}$$

where MLP is a lightweight projection network. This residual is injected into the intermediate temporal pipeline to provide unfiltered physical cues.

Overall, KEB strengthens the backbone by both adaptively modulating visual representations with motion saliency and preserving fine-grained kinematic deviations, thereby improving sensitivity to subtle execution errors.

2.3 Clinically-Guided Intent Branch

In addition to motion-enhanced features provided by the KEB, subtle errors can also arise from deviations relative to procedural intent. To capture this aspect, we introduce the Clinically-Guided Intent Branch (CIB), which provides explicit supervision based on standardized surgical knowledge.

We construct intent descriptions for predefined error categories from [21], reformulated as normative maneuver statements under clinical guidance, such as: *"Only one piercing per needle regrasp; no repeated piercings."*, *"Instruments remain fully visible at all times."*. These textual instructions are encoded using BioClinicalBERT [3], producing semantic embeddings $F_{text} \in \mathbb{R}^{1\times 768}$ that define the clinical reference space.

To align semantic intent with motion-enhanced execution features F_e, we concatenate the text embedding with a learnable query vector to form the intent query Q. The textual component encodes clinical norms, while the learnable query introduces adaptive visual grounding for precise probing of spatiotemporal features. Cross-attention is then applied to inject intent guidance into the execution features:

$$F_Z = F_e + \texttt{CrossAttn}(Q, F_e, F_e). \tag{3}$$

The CIB fuses motion-enhanced features with clinical intent to generate an intent-aligned representation F_Z, complementing KEB features and facilitating subtle error discrimination in subsequent temporal processing.

2.4 Intent-Execution Alignment

To further enhance intent-execution alignment and consolidate the discrimination boundary, we introduce an auxiliary Intent-Execution Alignment (IEA) module, inspired by contrastive alignment methods such as CLIP [17], to explicitly regulate the relationship between the kinematic residual F_k^{Δ} and the intent query Q.

Let $y_t \in \{0, 1\}$ denote the ground-truth label at frame t (0 for normal, 1 for error). Both $F_{k,t}^{\Delta}$ and Q_t are projected via MLP to a common 256-dimensional space. The IEA loss is defined asymmetrically as:

$$\mathcal{L}_{\text{IEA}}(F_{k,t}^{\Delta}, Q_t) = \begin{cases} \|F_{k,t}^{\Delta}\|_1, & y_t = 0 \\ \left|\texttt{cos}(F_{k,t}^{\Delta}, Q_t)\right|, & y_t = 1 \end{cases} \tag{4}$$

where $\texttt{cos}(\cdot,\cdot)$ denotes cosine similarity.

For normal maneuvers ($y_t = 0$), minimizing the L1 norm encourages $F^{\Delta}_{k,t} \to 0$, suppressing minor motion noise and maintaining stable execution aligned with intent. For error frames ($y_t = 1$), minimizing $|\texttt{cos}|$ forces the residual to diverge from the semantic intent, amplifying subtle discrepancies that are visually deceptive.

2.5 Loss Function

The final frame-level prediction combines intent-aligned features F_Z from CIB with kinematic residuals F^{Δ}_{k} from KEB. Both are first processed through $\texttt{MTemp}_0$ and summed, then passed through two temporal blocks and a classification MLP to produce $\hat{y}_t$. The overall loss jointly combines classification and alignment, using BCE loss and IEA loss, respectively:

$$\mathcal{L} = \sum_{t} (\mathcal{L}_{\text{BCE}}(y_t, \hat{y}_t) + \mathcal{L}_{\text{IEA}}(F^{\Delta}_{k,t}, Q_t)). \tag{5}$$

This joint supervision trains the network to classify subtle errors while enforcing alignment with semantic intent. BCE drives correct label prediction, while IEA leverages fine-grained kinematic cues to enlarge the margin between normal and erroneous executions.

3 Experiments

3.1 Experimental Design

Datasets. We evaluate KIDA on SAR-RARP50 [16] (sampled at 5 Hz), an in-vivo dataset containing suturing segments from Robot-Assisted Radical Prostatectomy (RARP) procedures. The fine-grained nature of suturing makes it suitable for evaluating subtle error detection. It comprises 40 training (46,586 normal and 18,540 error frames) and 8 testing videos (7,549 normal and 2,096 error frames). To validate generalizability, we additionally evaluate on the JIGSAWS benchmark [1] of elementary surgical tasks.

Metrics. We adopt standard frame-level evaluation metrics, namely area under the curve (AUC) and average precision (AP), to assess the detection performance. To evaluate sensitivity to transient faults, we further report results separately for short errors ($< 3s$) and long errors ($\geq 3s$).

Implementation Details. The proposed KIDA framework is trained end-to-end. All experiments are conducted on a single NVIDIA GeForce RTX 4090 24GB GPU. We use the AdamW optimizer with an initial learning rate of $2e-5$ and a weight decay of $1e-4$. The network is trained with a batch size of 1.

Table 1. Performance comparison on the SAR-RARP50 dataset. Fine-grained results are focusing the comparative analysis on highly competitive architectures.

Method	All Test Data		Short Errors ($< 3s$)		Long Errors ($\geq 3s$)	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
TeCNO [5]	58.14 ± 0.33	26.01 ± 0.32	-	-	-	-
MS-TCN [6]	58.82 ± 0.16	29.98 ± 3.06	-	-	-	-
MS-TCN++ [11]	58.57 ± 0.43	30.35 ± 1.35	-	-	-	-
LSTM [12]	68.43 ± 1.21	37.26 ± 2.65	60.00 ± 1.23	18.17 ± 1.57	79.25 ± 0.95	36.83 ± 1.68
ASFormer [22]	68.76 ± 0.93	36.79 ± 2.35	63.28 ± 0.77	25.13 ± 0.63	76.26 ± 0.42	33.61 ± 0.94
Vim [23]	69.06 ± 1.46	40.63 ± 2.30	62.57 ± 1.00	24.94 ± 0.28	77.18 ± 1.88	35.04 ± 4.48
Mamba [8]	69.38 ± 0.26	41.07 ± 0.95	63.17 ± 2.44	24.83 ± 3.67	74.17 ± 1.33	29.08 ± 1.41
SEDMamba [21]	71.20 ± 0.26	44.87 ± 1.52	62.24 ± 0.25	24.57 ± 0.21	79.57 ± 1.01	35.30 ± 1.52
Ours	$\mathbf{73.48 \pm 0.17}$	$\mathbf{49.73 \pm 0.29}$	$\mathbf{63.82 \pm 0.58}$	$\mathbf{26.96 \pm 1.42}$	$\mathbf{82.29 \pm 0.43}$	$\mathbf{44.06 \pm 1.80}$

3.2 Comparison with the State-of-the-arts

Quantitative and Qualitative Comparison. To evaluate the effectiveness of KIDA in detecting subtle surgical errors, we compare it with representative approaches, including video-based models (TeCNO [5], MS-TCN [6], MS-TCN++ [11]), Transformer-based methods (ASFormer [22], Vim [23]), temporal modeling networks (LSTM [12]), and recent selective state space models such as Mamba [8] and SEDMamba [21]. Experiments are first conducted on the widely used SAR-RARP50 benchmark, with quantitative results reported in Table 1.

Our method achieves state-of-the-art performance, surpassing previous SOTA method SEDMamba by 2.28% in AUC and 4.86% in AP, demonstrating superior discrimination of subtle execution-level errors. Consistent improvements are also observed for both long and short error categories. For prominent long errors ($\geq 3s$), AUC increases from 79.57% to 82.29%, indicating enhanced modeling of sustained deviations. For short errors ($< 3s$), although absolute gains are intrinsically constrained by temporal smoothing, AUC still improves from 62.24% to 63.82% and AP from 24.57% to 26.96%. These consistent improvements across all durations further confirm the effectiveness and practical relevance of our framework in resolving clinically critical subtle deviations.

We visualize the predicted error probability curves of our method and SEDMamba and present two representative cases in Fig. 2. As shown, our method generates noticeably sharper and higher peaks than SEDMamba, effectively alleviating the baseline's insufficient response to subtle errors, particularly in the red boxed regions. Rather than overly smoothing the predictions, KIDA amplifies the relative margin between erroneous and normal states, which is consistent with the observed improvement in overall AP.

Generalization Performance. To validate generalizability, we evaluate KIDA on the dry-lab JIGSAWS dataset, an environment inherently simpler than complex clinical scenarios due to its standardized tasks. As shown in Table 2, COG is excluded from direct comparison for utilizing additional gesture information. Under a fair setup, KIDA consistently surpasses SEDMamba in both AUC and AP. These steady improvements across the highly complex SAR-

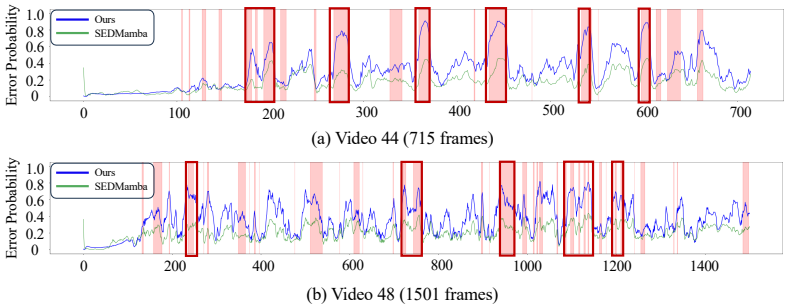

Fig. 2. Visualization of prediction results from the proposed KIDA (blue curve) and the previous state-of-the-art method SEDMamba (green curve). The pink shaded regions indicate ground-truth error frames. The horizontal axis represents the frame index, and the vertical axis denotes the predicted error probability. The red boxed areas highlight subtle errors that are missed by SEDMamba but successfully detected by KIDA.

Table 2. Performance comparison on the JIGSAWS dataset. COG is excluded from comparison as it utilizes additional gesture information.

Method	AUC (%)	AP (%)
COG [18]	72.92 ± 2.87	75.02 ± 6.25
SEDMamba [21]	72.69 ± 2.67	74.16 ± 6.01
Ours	**73.53 ± 2.88**	**74.73 ± 6.16**

Table 3. Ablation study of different components.

Method	AUC (%)	AP (%)
KIDA	**73.70 ± 0.34**	**49.48 ± 0.31**
w/o IEA	73.39 ± 0.52	48.43 ± 1.78
w/o IEA & KEB	72.10 ± 1.02	46.72 ± 0.93
w/o IEA & CIB	71.51 ± 0.25	45.47 ± 0.81
w/o IEA & KEB & CIB	71.20 ± 0.26	44.87 ± 1.52

RARP50 and the simplified JIGSAWS datasets confirm that explicit kinematic-intent alignment maintains generalizability across varying surgical complexities.

3.3 Ablation Study

Table 3 reports a step-wise ablation study of our framework. The full model KIDA achieves the best performance overall. Removing Intent-Execution Alignment (IEA) leads to drop in AP (Row 2), indicating that explicit intent-execution alignment helps produce more distinguishable error responses. Further removing Kinematic Execution Branch (KEB) causes a larger degradation (Row 3), showing that motion modeling is crucial for resolving visual ambiguity. Eliminating Clinically-Guided Intent Branch (CIB) reduces the model to an implicit visual baseline and yields the lowest performance (Row 4), highlighting the importance of clinically grounded semantic guidance.

We further replace clinical intents with irrelevant text and observe an AUC drop to 72.06%, indicating that clinically relevant intents provide effective task-specific semantic guidance beyond the mere presence of text.

Overall, the consistent decline across variants demonstrates that all three modules contribute complementary benefits, and their integration produces the strongest discrimination capability.

4 Conclusion

In this paper, we propose the Kinematic-Intent Dual-path Alignment (KIDA) framework, reformulating subtle surgical error detection as an explicit measurement of intent-execution consistency. By aligning physical execution with clinical references, our method robustly discriminates minor deviations. Experiments demonstrate state-of-the-art SAR-RARP50 performance and generalizability of JIGSAWS. Despite stable convergence, KIDA occasionally over-smooths dense errors and is limited to single-stage validation and dataset-constrained binary detection. Future work will explore finer-grained annotations for multi-class warning systems across broader procedures.

Acknowledgments. This work was supported by National Key R&D Program of China (Grant No. 2023YFC2414900), National Natural Science Foundation of China (Grant No. 62576145), and Wuhan United Imaging Healthcare Surgical Technology Co., Ltd.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Transactions on Biomedical Engineering **64**(9), 2025–2041 (2017)
2. Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z., Iyer, R.K.: Adverse events in robotic surgery: a retrospective study of 14 years of fda data. PloS one **11**(4), e0151470 (2016)
3. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W., Jin, D., Naumann, T., McDermott, M.B.A.: Publicly available clinical BERT embeddings. CoRR **abs/1904.03323** (2019), http://arxiv.org/abs/1904.03323
4. Chen, J.K., Chang, Y.J., Lin, C.B., Pan, Y., Wang, P.F.: Occurrence and impact of intraoperative anastomotic leakage in retzius-sparing robot-assisted radical prostatectomy. Medicina **61**(5), 886 (2025)
5. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International conference on medical image computing and computer-assisted intervention. pp. 343–352. Springer (2020)
6. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3575–3584 (2019)
7. Ghamrawi, W., Curtis, N., Xu, J., Boal, M., Mazomenos, E., Edwards, E., Stoyanov, D., Francis, N.: A rectal cancer surgery dataset: Use of artificial intelligence to aid automation of error identification. Scientific Data **12**(1), 180 (2025)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)

9. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
10. Hutchinson, K., Li, Z., Cantrell, L.A., Schenkman, N.S., Alemzadeh, H.: Analysis of executional and procedural errors in dry-lab robotic surgery experiments. The International Journal of Medical Robotics and Computer Assisted Surgery **18**(3), e2375 (2022)
11. Li, S.J., AbuFarha, Y., Liu, Y., Cheng, M.M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2020). https://doi.org/10.1109/TPAMI.2020.3021756
12. Li, Z., Hutchinson, K., Alemzadeh, H.: Runtime detection of executional errors in robot-assisted surgery. In: 2022 International conference on robotics and automation (ICRA). pp. 3850–3856. IEEE (2022)
13. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9517–9526 (2021). https://doi.org/10.1109/CVPR46437.2021.00940
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
15. Pompe, R.S., Tilki, D.: Complications after salvage radical prostatectomy: vesicourethral anastomosis leaks and possible prevention. Translational Andrology and Urology **6**(5), 994 (2017)
16. Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496 (2023)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Shao, Z., Xu, J., Stoyanov, D., Mazomenos, E.B., Jin, Y.: Think step by step: Chain-of-gesture prompting for error detection in robotic surgical videos. IEEE Robotics and Automation Letters (2024)
19. Sirajudeen, N., Boal, M., Anastasiou, D., Xu, J., Stoyanov, D., Kelly, J., Collins, J., Sridhar, A., Mazomenos, E., Francis, N.: Deep learning prediction of error and skill in robotic prostatectomy suturing. Surgical Endoscopy **38**(12), 7663–7671 (2024)
20. Vedula, S.S., Ishii, M., Hager, G.D.: Objective assessment of surgical technical skill and competency in the operating room. Annual review of biomedical engineering **19**, 301–325 (2017)
21. Xu, J., Sirajudeen, N., Boal, M., Francis, N., Stoyanov, D., Mazomenos, E.B.: Sedmamba: Enhancing selective state space modelling with bottleneck mechanism and fine-to-coarse temporal fusion for efficient error detection in robot-assisted surgery. IEEE Robotics and Automation Letters (2024)
22. Yi, F., Wen, H., Jiang, T.: Asformer: Transformer for action segmentation. arXiv preprint arXiv:2110.08568 (2021)
23. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)

24. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Ploetz, T., Clements, M.A., Essa, I.: Automated video-based assessment of surgical skills for training and evaluation in medical schools. International journal of computer assisted radiology and surgery **11**(9), 1623–1636 (2016)