R-MSA: Rhythm-Conditioned Surgical Video Skill Assessment via Morphology Synchronization and Uncertainty Calibration

Xiao Ke[1,2†], Yining Zheng[1,2†], and Huangbiao Xu[1,2⋆]

[1] Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[2] Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350108, China
kex@fzu.edu.cn, huangbiaoxu.chn@gmail.com

Abstract. Automated video-based surgical skill assessment enables objective feedback for training, especially in resource-limited settings. However, most existing methods rely on generic spatiotemporal modeling (e.g., absolute positional encoding or simple temporal aggregation) and overlook large execution rhythm variability across surgeons. For instance, experts tend to operate smoothly and efficiently, whereas novices exhibit slower and irregular patterns, which ultimately limits skill discrimination. To address this, we propose **R-MSA** (**R**hythm–**M**orphology **S**ynchronous **A**ssessment), a rhythm-aware framework that explicitly inherent models execution rhythm differences. Specifically, a Rhythm Morphology Encoder performs duration-conditioned rhythm encoding to generate rhythm-adaptive representations for variable-length videos. A Morphology-Synchronous Decoder with learnable queries aligns latent morphology primitives (e.g., needle driving, knot tying) across rhythm-normalized sequences. To refine procedural correspondence, a Surgical Sequence Alignment Module integrates a Gaussian temporal prior into co-attention for localized temporal alignment. Finally, an Uncertainty-Calibrated Interpretable Tree produces relative skill scores with calibrated uncertainty through reparameterized prediction. Experiments on JIGSAWS, HeiChole, and our proposed Kangduo datasets show that R-MSA consistently outperforms state-of-the-art methods in ranking correlation and error metrics. Our code is available at: https://bit.ly/R-MSA.

Keywords: Surgical skill assessment · Rhythm-aware video modeling · Duration-conditioned encoding · Morphology synchronization.

1 Introduction

Automated surgical skill assessment (SSA) is fundamental to computer-assisted surgical training, providing objective feedback to improve training efficiency and

⋆ Corresponding author.
† Equal contribution.

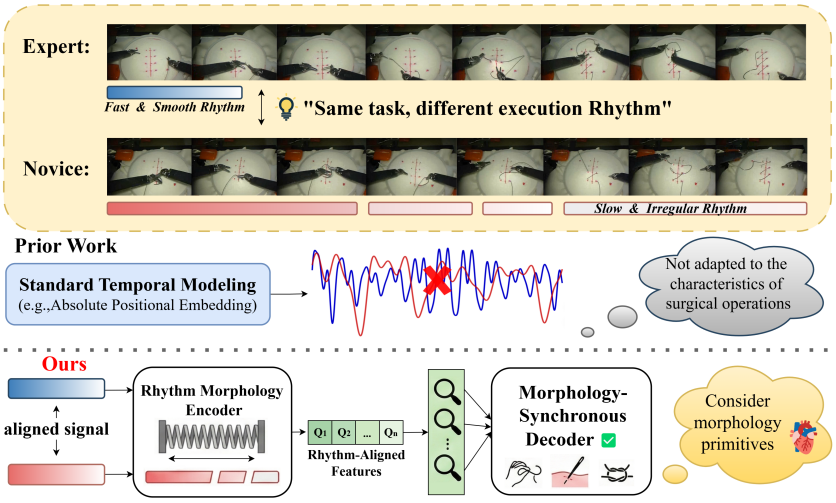

Fig. 1: Comparison of execution rhythms between expert and novice surgeons performing the same task. Experts exhibit fast and smooth rhythms, whereas novices show slow and irregular patterns. Such rhythm variability is a key skill indicator but is largely overlooked in existing temporal modeling approaches.

surgical safety [24,9,17]. Traditional expert scoring, however, is subjective, labor-intensive, and poorly reproducible [14,5]. The rapid adoption of robotic-assisted and minimally invasive surgery has produced large-scale surgical video data, enabling learning-based SSA. Nevertheless, two fundamental challenges remain [29]: capturing subtle differences in fine-grained surgical actions and modeling long-range temporal dependencies under substantial execution variability.

Recent SSA research has evolved from CNN-RNN pipelines to attention- and Transformer-based temporal modeling, including phase-aware attention [28], feature-score alignment [4], and event-aware frameworks such as SEDSkill [3]. Despite these advances, most approaches primarily model *what* actions occur while under-modeling *how* they unfold temporally across performers, particularly the interaction between execution rhythm and temporal morphology behavior [15].

Expert assessment relies on temporal motion evolution rather than isolated frames, reflecting control, precision, and fluency, termed *temporal morphology behavior* [15,10,32]. Stable trajectories indicate depth perception, prolonged dwells imply inefficiency, and abrupt motions suggest tissue risk [16]. These shape-invariant *morphology primitives* capture the essence of expertise. However, their manifestation in feature space is modulated by execution rhythm, the temporal arrangement and regularity of primitives within a procedure. Such modulation introduces distributional shifts, hindering fair comparisons across surgeons and durations. In this work, we formally define surgical rhythm as the temporal modulation pattern governing the arrangement and regularity of morphology primitives, rather than absolute execution speed or operation duration.

To address this, we propose **R-MSA** (Rhythm-Morphology Synchronous Assessment), a unified framework that models SSA as a progressive process of rhythm conditioning, sequence alignment, morphology synchronization, and uncertainty-aware regression. Specifically, the proposed Rhythm Morphology Encoder (RME) conditions temporal representations on execution rhythm, enabling rhythm-invariant encoding of morphology primitives. A Gaussian temporal prior is integrated into co-attention for localized sequence alignment. On rhythm-normalized features, a morphology-synchronous decoder employs learnable queries to extract structured primitives. Finally, a variational regression tree performs uncertainty-calibrated ranking to yield interpretable skill scores with reliable confidence intervals for clinical evaluation.

Furthermore, widely used benchmarks such as JIGSAWS [7] were collected a decade ago with low-resolution cameras and constrained environments, lacking the substantial execution variance needed to evaluate modern rhythm-aware models. To bridge this gap, we introduce the **Kangduo dataset**, comprising 104 high-definition robotic suturing videos. Characterized by complex visual details and significant execution rhythm variance, Kangduo provides a challenging and clinically realistic benchmark for surgical skill assessment.

Contributions. (1) We propose R-MSA, a rhythm-aware SSA framework that models execution rhythm via duration-conditioned encoding and morphology-synchronous decoding for robust comparison across heterogeneous rhythms. **(2)** We introduce physiology-inspired morphology primitives combined with a Gaussian temporal prior and variational regression tree for interpretable assessment with uncertainty quantification. **(3)** We present the Kangduo dataset (104 HD suturing videos with expert scores) as a challenging SSA benchmark. **(4)** Experiments on JIGSAWS [7], HeiChole [27], and Kangduo show consistent improvements over state-of-the-art methods in Spearman correlation and MAE.

2 Method

Figure 2 illustrates our **R-MSA** framework, which regresses relative surgical skill scores from video pairs. It comprises four modules: (a) Surgical Sequence Alignment Module (SSAM) for inter-video alignment, (b) Rhythm Morphology Encoder (RME) to normalize execution rhythms, (c) Morphology-Synchronous Decoder (MSD) to extract multi-level morphology primitives, and (d) Uncertainty-Calibrated Interpretable Tree (UCIT) for uncertainty-aware clinical scoring.

Feature Extraction. For fair comparison with prior methods [25,12,21,30], input videos are uniformly sampled into T clips and processed by a pre-trained I3D backbone, yielding spatiotemporal features $X \in \mathbb{R}^{T\times C}$ ($C = 1024$).

2.1 Surgical Sequence Alignment Module (SSAM)

Direct temporal comparison between surgical videos is challenging due to large duration and phase variations. To enable fine-grained cross-video alignment, we adopt SSAM that performs sample-wise co-attention with Gaussian temporal

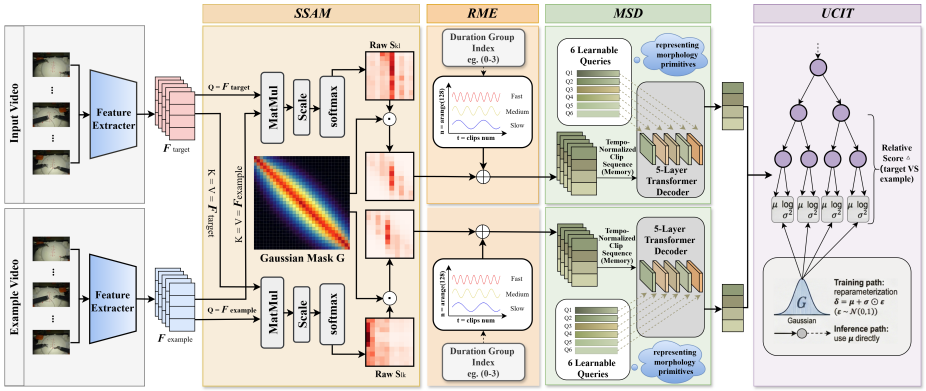

Fig. 2: Illustration of our R-MSA framework, which comprises SSAM, RME, MSD, and UCIT, forming a rhythm-aware morphology synchronization pipeline.

prior. The raw target video V_k and expert exemplar V_l are adaptively sampled into T clips ($T \leq T_{\max}$) and processed by the I3D backbone to obtain features $F_k, F_l \in \mathbb{R}^{N\times T\times C}$, where N represents the batch size. Following [26], we compute scaled dot-product attention independently for each sample b:

$$S_{kl}^{(b)} = \mathrm{softmax}\left(\frac{F_k^{(b)}(F_l^{(b)})^T}{\sqrt{C}}\right), \quad S_{lk}^{(b)} = \mathrm{softmax}\left(\frac{F_l^{(b)}(F_k^{(b)})^T}{\sqrt{C}}\right). \tag{1}$$

To suppress mismatched long-range interactions caused by rhythm variance, we introduce a Gaussian temporal prior mask $G \in \mathbb{R}^{T\times T}$:

$$G(i,j) = \exp\left(-\frac{(i-j)^2}{2(\gamma T)^2}\right), \quad \gamma = 0.2. \tag{2}$$

This prior reflects the clinical observation that comparable procedural phases tend to appear in temporally neighboring segments after adaptive sampling. The masked attention is then:

$$\tilde{S}_{kl} = S_{kl} \odot G, \quad \tilde{S}_{lk} = S_{lk} \odot G^T, \tag{3}$$

yielding aligned features $F_k^{\mathrm{enh}} = \tilde{S}_{lk}F_k$ and $F_l^{\mathrm{enh}} = \tilde{S}_{kl}F_l$.

2.2 Rhythm Morphology Encoder (RME)

Surgical videos of the same task exhibit large duration variance due to differing execution rhythms. To represent identical morphology patterns under different speeds, we introduce a rhythm-conditioned encoding instead of absolute positional encoding. Videos are divided into 4 quantile bins based on their sorted execution durations, producing a rhythm index $g \in \{0,1,2,3\}$. For position t and dimension n, the frequency is conditioned as:

$$\omega_{g,n} = \frac{g+1}{10000^{2n/D}}, \quad D = 128. \tag{4}$$

The rhythm encoding is computed by the sinusoidal positional encoding [26]:

$$R_{t,2n} = \sin(\omega_{g,n} t), \quad R_{t,2n+1} = \cos(\omega_{g,n} t). \tag{5}$$

This conditioning on video duration enables rhythm-invariant morphology modeling. The encoding is concatenated with enhanced features to produce rhythm-normalized representations $F^{\text{enh+rhy}} \in \mathbb{R}^{N \times T \times 1152}$.

2.3 Morphology-Synchronous Decoder (MSD)

We design a Morphology-Synchronous Decoder (MSD) to extract surgical morphology primitives invariant to varying durations and execution rhythms. The rhythm-normalized features are transposed into memory $M = (F^{\text{enh+rhy}})^T \in \mathbb{R}^{T \times N \times 1152}$. Six learnable queries $Q \in \mathbb{R}^{N \times K \times 1152}$ ($K = 6$) attend to this memory through a 5-layer transformer decoder (8 heads, $d_{\text{model}} = 1152$, FFN dim 2048, dropout 0.1, LayerNorm, positional encodings following [26]). The decoder employs cross-attention from queries to memory and self-attention among queries, yielding output $D \in \mathbb{R}^{N \times K \times 1152}$, where each query learns a distinct morphology primitive.

The number $K = 6$ is chosen as the optimal set that balances coverage of recurring patterns and overfitting risk (ablation shows $K = 6$ outperforms $K = 3$ and $K = 9$ by 2% and 3% in Spearman ρ; see Table 3). Video-level embeddings are obtained by mean-pooling over the query dimension:

$$F^{\text{video}} = \frac{1}{K} \sum_{i=1}^{K} D_i \in \mathbb{R}^{N \times 1152}. \tag{6}$$

This yields synchronized representations focused on action morphology.

2.4 Uncertainty-Calibrated Interpretable Tree (UCIT)

The concatenated target-exemplar video embeddings are fed into an Uncertainty-Calibrated Interpretable Tree (UCIT)—a differentiable soft decision tree (depth 5, 32 leaves) employing probabilistic routing and leaf-level Gaussian predictors. Internal nodes output branch probabilities $p_n = \sigma(w_n^T x)$, and the leaf routing probability p_l is the path product. Each leaf predicts a mean μ_l and log-variance $\log \sigma_l^2$, yielding the mixture prediction:

$$\mu = \sum_l p_l \mu_l, \quad \sigma^2 = \sum_l p_l \sigma_l^2. \tag{7}$$

During training, relative skill scores y (target minus exemplar ground-truth) are sampled via the reparameterization trick [13]. The overall objective $\mathcal{L} = \mathcal{L}_{\text{ranking}} + \lambda \mathcal{L}_{\text{uncertainty}}$ ($\lambda = 1$) incorporates a variance-attenuated loss:

$$\mathcal{L}_{\text{uncertainty}} = \alpha \frac{(y - \mu)^2}{\sigma^2} + \beta \log \sigma^2, \tag{8}$$

Table 1: Baseline comparison on JIGSAWS. We report the Spearmans Rank Correlations and MAE by three cross-validation schemes on every task. K: tic data, V: video frames, *: extra annotations such as surgical gestures are utilized.

Input	Method	KT			NP			SU			Avg		
		LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold
Spearman's Rank Correlation (SROCC)↑													
K+V	MultiPath-VTP [19]	-	0.58	0.78	-	0.62	0.76	-	0.45	0.79	-	0.55	0.78
	MultiPath-VTPE [19]	-	0.59	0.82	-	0.65	0.76	-	0.45	0.83	-	0.57	0.81
V	T^2CR [12]	-	-	0.89	-	-	0.89	-	-	0.93	-	-	0.91
	RICA^2 [21]	-	-	**0.90**	-	-	**0.94**	-	-	0.92	-	-	0.92
	ViSA [18]	0.92	0.76	0.84	0.93	0.90	0.86	0.84	0.72	0.79	0.90	0.81	0.83
	Contra-Sformer [1]	0.89	0.69	0.87	0.71	0.71	0.81	0.86	0.65	0.69	0.83	0.68	0.81
	MT-ViT [8]	**0.93**	0.85	0.89	0.86	0.88	0.82	0.87	0.75	0.82	0.89	0.83	0.84
	Ours	**0.93**	**0.94**	**0.90**	**0.89**	**0.95**	**0.94**	**0.96**	**0.92**	**0.95**	**0.92**	**0.94**	**0.93**
Mean Absolute Error (MAE)↓													
V	ViSA [18]	2.16	2.01	2.60	1.66	2.16	2.05	2.58	2.82	2.70	2.13	2.33	2.45
	Contra-Sformer [1]	1.75	**1.39**	2.10	3.15	3.17	3.21	2.74	2.58	2.99	2.55	2.38	2.77
	MT-ViT [8]	**1.69**	1.45	2.00	1.95	2.04	2.23	2.23	2.46	2.43	1.95	1.98	2.22
	Ours	1.90	1.94	**1.97**	**1.75**	**1.79**	**1.81**	**2.15**	**1.75**	**2.24**	**1.93**	**1.83**	**2.01**

where $\alpha = 0.6$ and $\beta = 0.4$. The first term adaptively attenuates penalties for ambiguous pairs via a larger σ^2, while the second regularizes against trivial variance inflation. At inference, following the validated exemplar-based paradigms of T^2CR [12], absolute skill is estimated by adding the predicted relative difference μ to the exemplar's ground-truth score.

3 Experiments

Dataset. We evaluate R-MSA on three surgical skill assessment datasets: JIGSAWS [7], HeiChole [27], and our proposed Kangduo dataset. JIGSAWS includes three robotic tasks (KT, NP, SU) with over 30 trials each on the da Vinci system. We use the sum score as ground truth [19,29,36] and report results under Leave-One-Supertrial-Out (LOSO), Leave-One-User-Out (LOUO), and 4-Fold cross-validation [22,6,25]. HeiChole comprises 24 endoscopic videos of laparoscopic cholecystectomy, with two clips from critical phases annotated across five domains. We use the sum score as ground truth and adopt 4-fold cross-validation with 75/25 split [29,36]. Kangduo is our high-definition robotic suturing dataset comprising 104 realistic training videos. Unlike JIGSAWS, it exhibits significant execution duration variance (14 mins), enabling robust rhythm-aware evaluation. Data collection followed standard ethical protocols (anonymized, IRB-exempted). To capture procedural nuances, ground truth scores were independently annotated by 3 experts using a tailored, institution-specific rubric to ensure high reliability. We use a 75/25 train/test split.

Evaluation Metrics. For JIGSAWS, we report Spearman's Rank Correlation Coefficient (SROCC) and Mean Absolute Error (MAE), with SROCC averaged via Fisher's z-value transformation [23,31,35,34] for baseline comparison. MAE is used in ablation studies. On HeiChole and Kangduo, SROCC and MAE are averaged over multiple runs with different random seeds for stability.

Implementation Details. We use Kinetics-pretrained I3D [2,11] to extract L2-normalized features from its last Inception block (earlier layers frozen, final fine-

Table 2: Results on HeiChole and Kangduo. R-MSA is compared with representative surgical skill assessment methods using MAE and SRCC.

Dataset	Method	MAE↓	SRCC↑
HeiChole Dataset	MUSDL [25]	1.58	0.23
	ViSA [18]	1.27	0.46
	MT-ViT [8]	1.06	0.58
	QGVL [33]	1.31	0.66
	ASGTN [20]	1.29	0.68
	Ours	**0.95**	**0.76**
Kangduo Dataset	MUSDL [25]	1.26	0.67
	ViSA [18]	1.19	0.75
	QGVL [33]	1.16	0.83
	ASGTN [20]	1.14	0.84
	Ours	**0.97**	**0.88**

Table 3: Ablation studies on JIGSAWS using 4-fold cross-validation.(a) Contribution of individual modules in R-MSA. (b) Influence of the number of morphology queries Q in the decoder.

(a) Module ablation

Model Variant	MAE↓	SRCC↑
Full (Ours)	**2.01**	**0.93**
w/o UCIT	2.16	0.91
w/o MSD	2.55	0.89
w/o RME	2.73	0.86
w/o SSAM	3.01	0.84

(b) Query number (Q)

Model Variant	MAE↓	SRCC↑
Full ($Q = 3$)	2.23	0.92
Full ($Q = 6$)	**2.01**	**0.93**
Full ($Q = 9$)	2.31	0.90

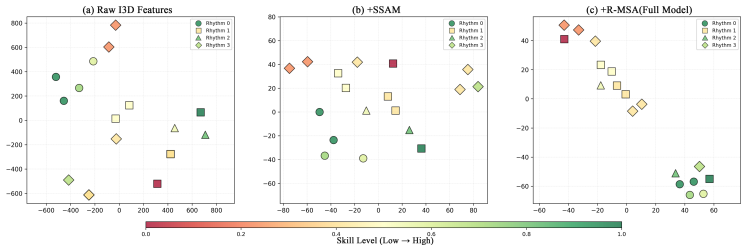

Fig. 3: Illustration of how rhythm-aware modeling progressively reshapes embedding geometry toward skill-aligned representations. Raw I3D features exhibit weak skill separability due to dominant rhythm variance. SSAM provides limited improvement, whereas the full R-MSA reorganizes the original high-dimensional space into clearly structured, skill-consistent clusters.

tuned). Videos are adaptively sampled into T clips (1 clip/s, capped at $T_{\max} = 20$). Frames are resized to 224×224 and ImageNet-normalized. Our PyTorch framework is trained on an RTX 3090 GPU for 200 epochs using AdamW (lr 1×10^{-4}, cosine decay, batch size 4).

3.1 Comparison with the State-of-the-Art Methods

Table 1 and Table 2 compare R-MSA with state-of-the-art methods on JIGSAWS, HeiChole, and Kangduo. On JIGSAWS, R-MSA outperforms visual-based methods (ViSA, Contra-Stomer, MT-ViT) and multi-modal approaches (MultiPath-VTPE) without kinematic data or extra annotations, achieving the best average SROCC (0.92-0.94) and lowest MAE (1.93-2.01) across all schemes. It also surpasses recent AQA methods T^2CR and RICA in overall robustness. On more

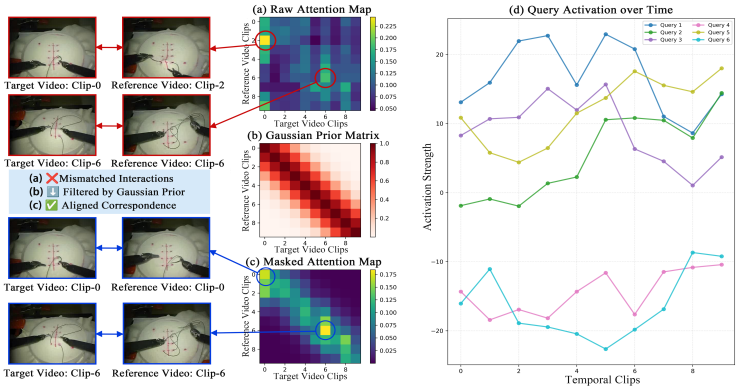

Fig. 4: Morphology Synchronization Mechanism. (a) Raw attention exhibits mismatched long-range interactions. (b) The Gaussian prior imposes temporal locality bias. (c) The masked attention focuses on aligned regions, highlighting target-reference correspondence. (d) Query activations over time reveal distinct morphology primitives.

challenging datasets, R-MSA attains the lowest MAE (HeiChole 0.95, Kangduo 0.97) and highest SROCC (HeiChole 0.76 Kangduo 0.88), significantly outperforming ViSA, MT-ViT, and ASGTN.

3.2 Ablation Study

Quantitative analysis. Table 3 details ablation results on JIGSAWS (4-fold cross-validation). Removing SSAM degrades performance, validating rhythm-aware feature alignment. Omitting RME further drops metrics, proving explicit rhythm encoding mitigates duration-induced distortion. Without MSD, impaired morphology primitive extraction noticeably declines both SROCC and MAE. Dropping UCIT reduces performance, confirming the value of cross-instance temporal calibration. Overall, all modules provide complementary benefits.

Query number analysis. Increasing morphology queries Q from 3 to 6 improves performance by enhancing representation capacity. However, larger Q yields no further gains risking over-fragmentation. The optimal $Q = 6$ suggests a balanced decomposition of surgical motions into distinct primitives.

Qualitative analysis. To understand the learned representations, we visualize embedding geometry and attention patterns on the test set. As shown in Fig. 3, raw I3D features exhibit weak skill-level separability (Silhouette = 0.0275), indicating dominance of execution rhythm variance. Incorporating SSAM slightly improves clustering (Silhouette = 0.0922), while the full R-MSA reorganizes the original high-dimensional feature space into clearly structured skill-consistent clusters (Silhouette = 0.3340), demonstrating rhythm-invariant morphology modeling. Figure 4 further illustrates Gaussian prior-guided attention and query activations over time. The masked attention suppresses spurious cross-temporal

interactions while preserving structured alignment, and distinct queries show differentiated activation patterns corresponding to diverse morphology primitives.

4 Conclusion

In this paper, a novel framework called R-MSA is proposed to predict the skill score from surgical videos by synchronizing morphology primitives across heterogeneous execution rhythms. The framework can achieve competitive performance on three datasets: JIGSAWS, HeiChole, and Kangduo as well as support uncertainty-calibrated score prediction for clinical interpretability.

Acknowledgments This work was supported in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the National Natural Science Foundation of China under Grant 61972097 and U21A20472, in part by the Major Scientific Research Project for Technology Promotes Police under Grant 2025YZ040003, 2024YZ040001, in part by the Natural Science Foundation of Fujian Province under Grant 2025J01536.

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

References

1. 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 Autom. Lett. **8**(3), 1755–1762 (2023)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
3. Ding, X., Xu, X., Li, X.: Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In: MICCAI. pp. 35–45. Springer (2023)
4. Feghoul, K., Maia, D.S., Amrani, M.E., Daoudi, M., Amad, A.: Spatial-temporal graph transformer for surgical skill assessment in simulation sessions. In: Iberoamerican Congress on Pattern Recognition. pp. 287–297. Springer (2023)
5. Funke, I., Mees, S.T., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3d convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1217–1225 (2019)
6. Gao, J., Zheng, W.S., Pan, J.H., Gao, C., Wang, Y., Zeng, W., Lai, J.: An asymmetric modeling for action assessment. In: ECCV. pp. 222–238. Springer (2020)
7. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI workshop. vol. 3, p. 3 (2014)
8. Guo, J., Han, S., Liu, Y.H.: Mt-vit: Multi-task video transformer for surgical skill assessment from streaming long-term video. In: ROBIO. pp. 1752–1757. IEEE (2024)

9. Haidegger, T., Speidel, S., Stoyanov, D., Satava, R.M.: Robot-assisted minimally invasive surgerysurgical robotics in the data age. Proceedings of the IEEE **110**(7), 835–846 (2022)
10. Hutchinson, K., Chen, K., Alemzadeh, H.: Towards interpretable motion-level skill assessment in robotic surgery. arXiv preprint arXiv:2311.05838 (2023)
11. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
12. Ke, X., Xu, H., Lin, X., Guo, W.: Two-path target-aware contrastive regression for action quality assessment. Inf. Sci. **664**, 120347 (2024)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Lavanchy, J.L., Zindel, J., Kirtac, K., Twick, I., Hosgor, E., Candinas, D., Beldi, G.: Automation of surgical skill assessment using a three-stage machine learning algorithm. Scientific reports **11**(1), 5197 (2021)
15. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: CVPR. pp. 156–165 (2017)
16. Lea, C., Vidal, R., Hager, G.D.: Learning convolutional action primitives for fine-grained action recognition. In: ICRA. pp. 1642–1649. IEEE (2016)
17. Li, Y., Zhao, Z., Li, R., Li, F.: Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends. Artif. Intell. Rev. **57**(11), 291 (2024)
18. Li, Z., Gu, L., Wang, W., Nakamura, R., Sato, Y.: Surgical skill assessment via video semantic aggregation. In: MICCAI. pp. 410–420. Springer (2022)
19. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: CVPR. pp. 9522–9531 (2021)
20. Liu, J., Wang, H., Zhou, W., Stawarz, K., Corcoran, P., Chen, Y., Liu, H.: Adaptive spatiotemporal graph transformer network for action quality assessment. IEEE Trans. Circuits Syst. Video Technol. (2025)
21. Majeedi, A., Gajjala, V.R., GNVV, S.S.S.N., Li, Y.: Rica 2: Rubric-informed, calibrated assessment of actions. In: ECCV. pp. 143–161 (2024)
22. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: ICCV. pp. 6331–6340 (2019)
23. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: WACV. pp. 1468–1476. IEEE (2019)
24. Rivas-Blanco, I., Perez-Del-Pulgar, C.J., Garcia-Morales, I., Munoz, V.F.: A review on deep learning in minimally invasive surgery. IEEE Access **9**, 48658–48678 (2021)
25. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment. In: CVPR. pp. 9839–9848 (2020)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
27. Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical Image Anal. **86**, 102770 (2023)
28. Wan, B., Peven, M., Hager, G., Sikder, S., Vedula, S.S.: Spatial-temporal attention for video-based assessment of intraoperative surgical skill. Scientific reports **14**(1), 26912 (2024)
29. Wang, T., Wang, Y., Li, M.: Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In: MICCAI. pp. 668–678. Springer (2020)

30. Xu, H., Ke, X., Li, Y., Xu, R., Wu, H., Lin, X., Guo, W.: Vision-language action knowledge learning for semantic-aware action quality assessment. In: ECCV. pp. 423–440 (2024)
31. Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Guo, W.: Language-guided audio-visual learning for long-term sports assessment. In: CVPR. pp. 23967–23977 (2025)
32. Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Xu, P., Guo, W.: Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In: AAAI. pp. 8869–8877 (2025)
33. Xu, H., Wu, H., Ke, X., Li, Y., Xu, R., Guo, W.: Quality-guided vision-language learning for long-term action quality assessment. IEEE Trans. Multim. **27**, 7326–7339 (2025)
34. Xu, H., Wu, H., Ke, X., Peng, Y.: Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. arXiv preprint arXiv:2605.00434 (2026)
35. Xu, H., Wu, H., Ke, X., Wu, J., Xu, R., Xu, J.: Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In: AAAI. pp. 11241–11249 (2026)
36. Zia, A., Essa, I.: Automated surgical skill assessment in rmis training. Int. J. Comput. Assist. Radiol. Surg. **13**(5), 731–739 (2018)