This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Can Vision-Language Models Safely Rate Robotic Surgical Skill? A Failure Mode Analysis with Video, Kinematics, and Expert References

Siying Chen(✉) and Xianyi Cheng

Duke University, Durham, NC, USA
{siying.chen, xianyi.cheng}@duke.edu

Abstract. Expert assessment of robotic surgical skills provides feedback that trainees need but is expensive to scale, and wrong automated scores can mislead training. Whether general-purpose vision-language models (VLMs) can serve as reliable skill raters, and whether their errors stem from missing information or from failing to use the information they are given remains untested. We evaluate GPT-5.5, Claude Opus 4.8, and Gemini 3.5 Flash as zero-shot raters on JIGSAWS, which pairs surgical video and da Vinci kinematics with expert modified OSATS Global Rating Scale (GRS) scores. We find that VLMs can rank surgical performances from video but cannot score them at expert level: the best performances are under-scored, the worst are over-scored, and predictions cluster near the middle of the 6 to 30 scale. Even the best setting, GPT-5.5 at 3 frames per second, reaches a Spearman correlation of 0.58, indicating only moderate agreement with the expert ordering. However, the errors do not come from missing motion information. Given a text summary of the kinematics instead of video, the VLMs agree with experts near zero (ICC $\leq$ 0.06), yet a linear regressor fed with the same summary reaches ICC = 0.44; and adding kinematic summaries to the video lowers GPT-5.5 to $\rho = 0.39$. Finally, showing the model expert's score references spreads its predictions toward the expert score range, but does not improve the ranking. Overall, current VLMs can see useful surgical cues, but are not reliable standalone skill raters on dry-lab robotic tasks.

Keywords: Robotic surgical skill assessment · AI-assisted surgical video assessment · Vision-language models.

1 Introduction

Technical skill is a central component of surgical safety. Expert assessment gives trainees feedback not only on whether a task was completed, but also on how safely, efficiently, and reliably it was performed, and surgical skill has been linked to patient outcomes [2]. Structured tools such as OSATS and GEARS are widely

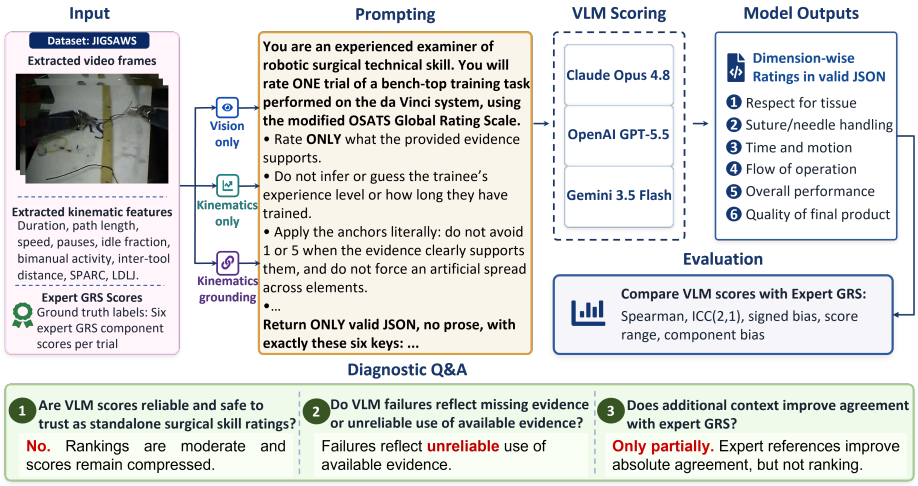


Fig. 1. Workflow for diagnosing VLM surgical skill scoring failures. For each JIGSAWS trial, the model receives one of three controlled inputs: sampled video frames only, extracted kinematic features only, or both. The VLM returns dimension-wise modified OSATS Global Rating Scale (GRS) ratings, which are compared with expert GRS scores to assess reliability, safety, and evidence integration.

used to evaluate operative performance in simulation and robotic surgery training [11, 7]. However, expert assessment is difficult to scale because it requires trained raters, substantial review time, and careful management of inter-rater variability. This is especially limiting in surgical education, where frequent feedback is needed but expert time is scarce.

Assessment is also challenging because technical performance is multidimensional and only partially observable [12]. Experts do not judge motion efficiency alone; they also consider tissue handling, needle and suture control, flow of operation, final-product quality, and the decisions made throughout the procedure. As a result, a trainee may move smoothly yet apply inappropriate tension, make poor procedural choices, or produce an inadequate final product. Surgical skill assessment is therefore not video classification, but a perception-and-reasoning problem that requires expert clinical judgment.

At the same time, automated skill assessment has advanced rapidly, with supervised models achieving high agreement with expert scores on datasets such as JIGSAWS [10, 5]. Yet safe use of AI for the assessment of surgical skills requires more than high average prediction accuracy. A model may rank trainees correctly but still assign misleading scores. Such errors are safety-relevant because trainees may use these scores to judge their own readiness and guide practice. Therefore, it remains unclear whether general-purpose vision-language models (VLMs) can serve as reliable surgical skill raters, and whether their failures come from missing evidence or from poor use of available evidence.

We follow the workflow in Fig. 1. For each JIGSAWS trial, we construct controlled inputs from sampled video frames and kinematic features, including vision-only, kinematics-only, and kinematic-grounding conditions. The prompt asks each VLM to rate the six modified OSATS Global Rating Scale (GRS) dimensions as JSON without access to surgeon identity, experience level, or expert ratings. This setup addresses three diagnostic questions: (i) whether VLM scores are reliable and safe as standalone ratings; (ii) whether failures arise from missing evidence or unreliable use of available evidence; and (iii) whether added context such as kinematic grounding or expert references improves agreement with experts.

We find that current VLMs can recover a partial skill ranking from surgical video, but this ability does not translate into scores that are safe to trust. Firstly, even in the best direct setting, predicted scores are compressed toward the middle of the scale, under-scoring strong trials and over-scoring weak ones, so a correct ranking does not imply calibrated or trustworthy ratings. Secondly, these failures are not explained by missing motion information: supplying kinematic summaries, or even feature-level explanations of them, does not reliably improve agreement and can degrade it, which points to weak evidence integration ability rather than an information gap. Finally, providing expert-scored references for the same-task improves absolute agreement with the expert scale but leaves this integration problem unresolved. Moreover, these behaviors are strongly model dependent, with GPT-5.5 agreeing best (Spearman $\rho = 0.58$) while Claude is weaker and Gemini is near chance, and the errors follow a systematic pattern, with models giving high scores to what video cannot show, such as respect for tissue, and low scores to what requires judging the full procedure and its outcome. Together, these results show that, on the tested JIGSAWS dry-lab tasks, the bottleneck is not access to evidence but the ability to turn it into calibrated scores, and that current VLMs are better positioned as decision support than as standalone raters.

2 Related Work

Surgical skill assessment has traditionally relied on expert observation and structured rating scales. OSATS evaluates technical performance through component scores such as tissue handling, instrument use, flow, and overall performance [11], while GEARS adapts this idea to robotic surgery with dimensions such as bimanual dexterity, efficiency, force sensitivity, and robotic control [7]. These tools make expert judgment more systematic, but they still require trained raters and careful review. Automated skill assessment has also been studied on JIGSAWS. Kinematic methods show that global motion descriptors and learned time-series representations can predict expert skill scores, suggesting that motion efficiency, speed, smoothness, pauses, and bimanual coordination contain skill-relevant information [20, 15, 9]. Methods based on video and multimodal inputs broaden this evidence beyond robot trajectories by using visual motion patterns, tool-tissue interactions, and procedural events [5, 10]. However, these

systems are usually trained in a supervised manner for a specific dataset or task. They show that surgical skill leaves measurable traces in kinematics and video, but they do not directly test whether general-purpose VLMs can apply an expert GRS rubric without task-specific training, or whether VLM failures come from missing evidence or poor use of available evidence.

More recent work has explored foundation models for surgical video understanding. Although these models improve recognition of surgical actions, workflows, and visual cues, benchmarks still show that fine-grained OSATS-style skill prediction remains difficult [16, 8, 18]. Specialized surgical VLMs and their evaluation benchmarks report that both domain specific and general models still fall short of clinical requirements on skill-related tasks [19, 3], and general-purpose VLMs applied to laparoscopic video at low frame rates perform well on object recognition but poorly on finer clinical judgments [14]. Using large language models as evaluators is also known to be unreliable, with scores sensitive to the rating scale rather than the content judged [4]. Retrieval augmented generation may help by providing external context at inference time [17], but for surgical skill assessment it remains unclear which context is useful: raw kinematics, feature explanations, or expert-scored examples. We therefore evaluate general-purpose VLMs mainly as zero-shot surgical skill raters, and separately test whether expert-scored reference examples in the prompt can reduce their scoring errors. Beyond prediction performance, we also analyze whether added context improves agreement with expert GRS ratings and which failure modes matter for safe feedback.

3 Methods

3.1 Dataset

We use JIGSAWS, a public robotic surgery training dataset collected on the da Vinci platform [6]. It contains synchronized surgical video, robot kinematics, self-reported experience level, and modified OSATS GRS scores for 103 labeled trials across three tasks. We exclude one unusually long trial (Suturing_G001, 300.9 s), which would exceed the 600 image Claude request limit at 2 and 3 fps. The resulting paired evaluation set contains 102 trials: Suturing (38), Knot Tying (36), and Needle Passing (28). Each included trial has six expert GRS component scores (respect for tissue, suture/needle handling, time and motion, flow of operation, overall performance, and quality of final product), each scored 1–5, giving a total GRS range of 6–30.

3.2 Evaluation Metrics

The primary outcome is agreement between model-predicted and expert total GRS. We report Spearman rank correlation to test whether predictions preserve the expert ordering, and ICC(2,1) (two-way random effects, absolute agreement [13]) to test whether the predicted scores match the expert scores in value.

Unlike Spearman ρ, ICC is sensitive to systematic offsets and range compression, which are central to our analysis. Also, we report 95% confidence intervals from 2,000 bootstrap resamples over trials, and key condition comparisons use paired bootstrap differences computed on the same resampled indices. Following our exclusion criterion, all conditions are evaluated on the same 102 trials.

3.3 VLM Skill Scoring Conditions

We use all VLMs as zero-shot raters, without task specific training or fine tuning. Each prompt contains only the task name, the modified OSATS GRS rubric, and one evidence condition, and is blinded to surgeon identity, experience level, expert scores, and study hypotheses. For each GRS component the model returns one evidence sentence and one integer score from 1 to 5 as structured JSON, and the predicted total GRS is the sum of the six component scores. In plain terms, the prompt instructs each model to act as an experienced examiner, to score only what the visible evidence supports without inferring the trainee's experience, and to apply the rubric anchors literally rather than forcing an artificial spread.

Model and frame rate study. To test whether VLM skill scoring is model dependent and limited by sparse visual observation, we varied two factors: model family and visual sampling density. We evaluated three frontier VLMs: GPT-5.5, Claude Opus 4.8, and Gemini 3.5 Flash. For each model, we sampled surgical video frames at 1, 2, and 3 fps to examine whether denser temporal evidence improves agreement with expert skill rankings. All models were accessed in June 2026; because proprietary versions change over time, we treat the resulting model ranking as a snapshot rather than a fixed result.

Evidence condition study. To test whether VLM failures are caused by missing motion information or by inability to use explicitly provided motion evidence, we compared three evidence conditions: vision-only, kinematics-only, and kinematics-grounding. In the vision-only condition, the model received only sampled frames. In the kinematics-only condition, the model received a text summary of patient side tool motion only. In the kinematics-grounding condition, the model received both frames and the kinematic summary. The kinematic summary included duration, path length, speed, pauses, idle fraction, bimanual activity, distance between tools, spectral arc length (SPARC), and logarithmic dimensionless jerk [1]. To test whether this summary carries usable skill signal, we fit a ridge regressor on the same nine features, using leave-one-surgeon-out cross-validation and the same pooled protocol as the VLM conditions. With no feature engineering and no tuning beyond one regularization grid, it serves as an information-matched floor, not a competitive baseline.

Additional context conditions. Furthermore, we test which type of additional context helps GPT use the GRS scale more effectively. We compare two context strategies. The first is interpreted grounding, where the model receives the same frames and motion summary as in kinematics-grounding, plus short explanations of each motion feature and within-task percentiles. The second is expert reference retrieval, where GPT receives a few same task reference examples with expert GRS scores. These references cover low (GRS $\leq$ 14), middle (15–21), and

Table 1. Agreement between VLM predictions and expert total GRS. We report Spearman ρ for ranking agreement and ICC(2,1) for absolute agreement. Kinematic grounding adds the target-trial kinematic summary to video; interpreted grounding additionally provides feature definitions and within-task percentiles; expert references provide same-task expert-scored examples; Ref. + kin. also includes the target-trial kinematics. †Conditions evaluated only for GPT-5.5 at 1 fps. A ridge regressor using the same kinematic summary reaches $\rho = 0.45$ and ICC = 0.44.

	GPT-5.5		Claude Opus 4.8		Gemini 3.5 Flash	
Condition	ρ	ICC	ρ	ICC	ρ	ICC
Kinematics-only	0.00	0.00	-0.01	0.00	0.16	0.06
Vision-only, 1 fps	0.47	0.24	0.32	0.22	-0.02	0.01
Vision-only, 2 fps	0.44	0.30	0.32	0.15	-0.08	-0.08
Vision-only, 3 fps	0.58	0.38	0.23	0.14	-0.09	-0.09
Kin. grounding, 1 fps	0.24	0.11	0.22	0.12	0.08	0.05
Kin. grounding, 2 fps	0.38	0.22	0.33	0.08	-0.18	-0.14
Kin. grounding, 3 fps	0.39	0.25	0.28	0.10	-0.05	-0.08
Interp. grounding, 1 fps†	0.37	0.23	–	–	–	–
Expert refs., 1 fps†	0.49	0.43	–	–	–	–
Ref. + kin., 1 fps†	0.43	0.36	–	–	–	–

high (≥ 22) score ranges, so the model can compare the target trial with examples from different parts of the GRS scale. We evaluate two retrieval variants: references only, and references plus the target trial's kinematic summary. As a prompt-design control, we additionally score a 30-trial stratified subset (10 low, 10 middle, 10 high expert GRS) at 1 fps under two relaxed prompt variants.

4 Results

4.1 VLMs Recover Partial Rankings but Remain Unsafe as Raters

Table 1 shows that, on the tested JIGSAWS dry-lab tasks, VLMs recover partial surgical skill rankings, but this ability is strongly model dependent: GPT-5.5 ranks moderately in line with experts, whereas Claude Opus 4.8 is weaker and Gemini 3.5 Flash is near chance. Even for GPT-5.5, agreement did not vary monotonically with frame rate: the three settings gave $\rho = 0.47$ (95% CI $[0.30, 0.62]$), 0.44 ($[0.28, 0.58]$), and 0.58 ($[0.41, 0.71]$) at 1/2/3 fps, with 2 fps scoring below 1 fps. Because these intervals overlap substantially, no single frame rate reliably outperforms another; we therefore report 3 fps as GPT-5.5's best-performing setting without claiming that denser sampling reliably helps.

More importantly, ranking agreement does not imply that the scores are safe to trust as standalone ratings. Even in the best setting, GPT-5.5's predictions spanned only 11–19 of the 6–30 GRS scale (SD 2.22 versus 5.62 for expert ratings), indicating substantial score compression. The task-stratified scatter plots in Fig. 2(b–d) show the same pattern: predictions preserve some within-task

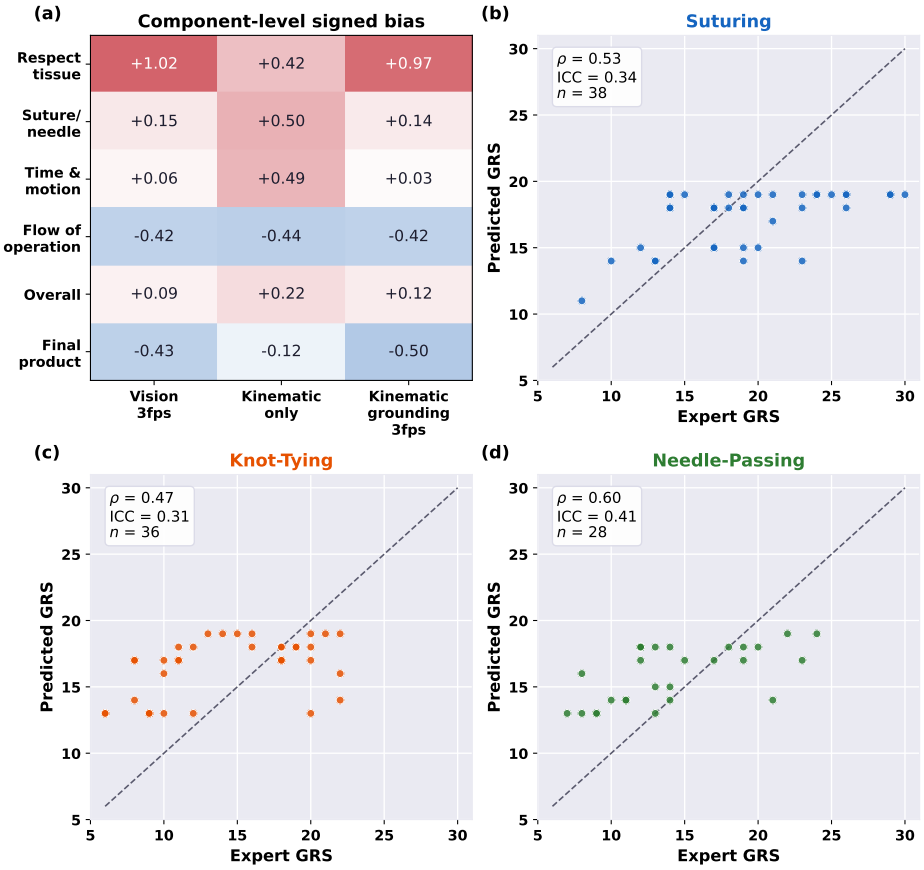


Fig. 2. GPT-5.5 failure modes. (a) Component-level signed bias, computed as predicted component score minus expert component score. Positive values indicate over scoring and negative values indicate under scoring. (b–d) Predicted versus expert total GRS for GPT-5.5 vision-only scoring at 3 fps, stratified by task. The dashed line denotes perfect agreement.

ordering, especially for Suturing and Needle Passing, but many points remain compressed toward the middle of the scale. High expert scores are often under predicted, while lower expert scores can be pushed upward into a seemingly acceptable range.

4.2 Failures Reflect Poor Evidence Integration, Not Missing Information

Before asking whether kinematic context helps the VLMs, we first test whether the kinematic summary carries usable skill signal at all. Under the same pooled protocol used for the VLMs, the ridge regressor fit to the identical kinematic

summary reaches $\rho = 0.45$ (95% CI $[0.27, 0.59]$) and ICC $= 0.44$ ($[0.28, 0.57]$), whereas all three VLMs given the same summary in the kinematics-only condition remain near zero (ICC ≤ 0.06). Thus, the kinematic summary is not uninformative; rather, the VLMs fail to use this structured motion evidence reliably under the tested prompting conditions. Kinematics-grounding likewise failed to help. For GPT-5.5 at 3 fps, adding kinematics of target trials significantly reduced agreement relative to vision-only, lowering Spearman ρ by 0.19 (paired bootstrap 95% CI $[0.02, 0.37]$) and ICC by 0.13 ($[0.01, 0.27]$), and kinematic grounding did not exceed vision-only scoring at any frame rate. Interpreted grounding partially recovered this loss (ICC 0.23 vs. 0.11 for kinematic grounding at 1 fps) but still fell short of the vision-only baseline (0.24), so explaining the features does not resolve the failure either.

Expert references provided a more useful form of context than raw motion summaries. As shown in Table 1, adding the same-task expert-scored examples at 1 fps raised the absolute agreement of GPT-5.5 with experts from ICC 0.24 to 0.43, while its ranking barely changed (ρ went from 0.47 to 0.49). This means references pull the predicted scores toward the range experts actually use, correcting the compression, but do not help the model tell better from worse trials. Adding target-trial kinematics on top of references reduced agreement again to $\rho = 0.43$, ICC $= 0.36$, which implies that references teach the model how experts use the GRS scale, but do not repair the underlying evidence-integration problem.

Beyond the aggregate GRS score, component-level analysis in figure 2(a) shows that GPT-5.5's errors are structured across GRS components rather than uniform. The largest bias is a strong positive offset on respect for tissue (+1.02 in the vision-only 3 fps condition), a dimension that depends on tissue tension and handling forces that are essentially unobservable from video; the model defaults to a favorable score in the absence of visible evidence. In contrast, the directly observable execution dimensions, time and motion (+0.06), overall performance (+0.09), and suture/needle handling (+0.15), are the best calibrated. The model then under-scores flow of operation (−0.42) and quality of final product (−0.43), the two dimensions that require integrating the whole procedure over time and judging the end result. In short, the model scores well on what the video shows and poorly on what it does not: it over-scores safety-critical dimensions it cannot see, such as respect for tissue, and under-scores those requiring judgment of the whole procedure and its outcome. This pattern was robust to prompt design: on the 30-trial ablation subset, both relaxed prompts produced the same component biases, so it is not an artifact of one prompt.

5 Conclusion and Future Work

In this work, we evaluated general-purpose VLMs as raters of robotic surgical skill under controlled evidence conditions. The diagnosis is consistent across experiments: VLMs recover partial skill rankings but compress scores, fail to exploit kinematic evidence that a trivial regressor extracts, and default to favor-

able ratings on safety-critical dimensions they cannot verify from video, while expert references improve calibration without repairing evidence integration. These findings suggest that, for dry-lab robotic tasks of the kind in JIGSAWS, VLM-based surgical skill assessment should be treated as decision support rather than autonomous scoring.

Our evidence is confined to JIGSAWS, a small dry-lab dataset, and to proprietary frontier models whose behavior may shift across versions. Evaluating open-weight models such as Qwen or InternVL is a natural next step, since their accessibility would enable reproducible analysis and fine-tuning. A further limitation is that we reduce skill to a single holistic GRS score, which omits the finer safety signals experts reason over. Assessing skill through intermediate cues that are directly visible in the video, such as whether the critical view of safety is achieved or how tissue deforms, offers a promising path that targets exactly the dimensions the models currently over-score.

Ethical approval. The original JIGSAWS data collection was approved by the Johns Hopkins University Institutional Review Board. This study performed a secondary analysis of de-identified dry-lab recordings and involved no new participant recruitment or data collection.

Declaration of Generative AI Use. The authors used OpenAI Codex to assist with drafting and language refinement. All scientific content, analyses, citations, and interpretations were reviewed and verified by the authors, who take full responsibility for the final manuscript.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Balasubramanian, S., Melendez-Calderon, A., Roby-Brami, A., Burdet, E.: On the analysis of movement smoothness. Journal of neuroengineering and rehabilitation **12**(1), 112 (2015)
2. Birkmeyer, J.D., Finks, J.F., O'reilly, A., Oerline, M., Carlin, A.M., Nunn, A.R., Dimick, J., Banerjee, M., Birkmeyer, N.J.: Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine **369**(15), 1434–1442 (2013)
3. Cheng, J., Zhao, X., Liu, S., Yu, X., Prakash, R., Codd, P.J., Katz, J.E., Lin, S.: Surgxbench: Explainable vision-language model benchmark for surgery. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8188–8198 (2026)
4. Fujinuma, Y.: Contrastive decoding mitigates score range bias in llm-as-a-judge. In: Findings of the Association for Computational Linguistics: ACL 2026. pp. 13404–13418 (2026)
5. 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**(7), 1217–1225 (2019)
6. 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: M2cai. vol. 3, p. 3 (2014)

7. Goh, A.C., Goldfarb, D.W., Sander, J.C., Miles, B.J., Dunkin, B.J.: Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. The Journal of urology **187**(1), 247–252 (2012)
8. Hoffmann, H., Bady, S., de Boer, C., Kirchner, M., Egger, J., Röhrig, R., Hölzle, F., Gruber, L.J., Xie, K., Neuhaus, M., et al.: Oss: Open suturing skills vision-based assessment challenge 2024-2025. arXiv preprint arXiv:2605.22200 (2026)
9. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. International journal of computer assisted radiology and surgery **14**(9), 1611–1617 (2019)
10. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9522–9531 (2021)
11. Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British journal of surgery **84**(2), 273–278 (1997)
12. Reiley, C.E., Lin, H.C., Yuh, D.D., Hager, G.D.: Review of methods for objective surgical skill evaluation. Surgical endoscopy **25**(2), 356–366 (2011)
13. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979)
14. Stueker, E.H., Kolbinger, F.R., Saldanha, O.L., Digomann, D., Pistorius, S., Oehme, F., Van Treeck, M., Ferber, D., Löffler, C.M.L., Weitz, J., et al.: Vision-language models for automated video analysis and documentation in laparoscopic surgery: a proof-of-concept study. International Journal of Surgery **111**(11), 7777–7786 (2025)
15. Wang, Z., Majewicz Fey, A.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. International journal of computer assisted radiology and surgery **13**(12), 1959–1970 (2018)
16. Wei, J., Xiao, Z., Sun, D., Gong, L., Yang, Z., Liu, Z., Wu, J.: Surgbench: A unified large-scale benchmark for surgical video analysis. arXiv preprint arXiv:2506.07603 (2025)
17. Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 6233–6251 (2024)
18. Xu, M., Shen, D., Zhang, J., Yip, H.C., Gao, Y., Chen, C., Imans, D., Long, Y., Ye, Y., Liu, Y., et al.: Generalized recognition of basic surgical actions enables skill assessment and vision-language-model-based surgical planning. arXiv preprint arXiv:2603.12787 (2026)
19. Zeng, Z., Zhuo, Z., Jia, X., Zhang, E., Wu, J., Zhang, J., Wang, Y., Low, C.H., Jiang, J., Zheng, Z., et al.: Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. arXiv preprint arXiv:2506.02555 (2025)
20. Zia, A., Essa, I.: Automated surgical skill assessment in rmis training. International journal of computer assisted radiology and surgery **13**(5), 731–739 (2018)