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.

SurgSkill-Bench: A Benchmark for Multimodal Surgical Skill Assessment

Chaohui Dang[1], Zheheng Jiang[2], James Glasbey[3], David Luke[4], Theodoros Arvanitis[1], and Le Zhang[1⋆]

[1] Department of Electronic, Electrical and Systems Engineering, School of Engineering, University of Birmingham, Birmingham, UK
[2] School of Computing and Mathematical Science, University of Leicester, Leicester, UK
[3] Department of Applied Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, UK
[4] University Hospitals of North Midlands NHS Trust, Stoke-on-Trent, UK
l.zhang.16@bham.ac.uk

Abstract. Objective assessment of surgical technical skill is important for surgical training and structured feedback, but current workflows remain dependent on labor-intensive expert review. Existing automated approaches primarily focus on visual inputs and provide limited support for jointly studying operative performance, structured skill scores, and evaluator feedback. We introduce SurgSkill-Bench, an initial video-score-text benchmark-style dataset containing 214 surgical training simulation videos, six-dimensional OSATS scores, and expert free-text comments. We define two evaluation settings: video-only OSATS prediction for automated assessment and post hoc expert-comment-assisted prediction, where evaluator comments are available as auxiliary information. We provide controlled baseline experiments using representative frozen visual backbones, content-adaptive key-frame sampling, and a simple video-text co-attention fusion module. Under internal video-level validation, content-adaptive sampling improves video-only performance in this dataset, while evaluator comments provide additional score-related signal in the assisted setting. The best mean AUROC reaches 0.88 under dataset-specific median dichotomization. We further discuss evaluation constraints related to dataset scale, metadata completeness, and the interpretation of comment-assisted prediction. *Code will be public after review.*

Keywords: surgical skill assessment · SurgSkill-Bench · multimodal learning · co-attention

1 Introduction

Objective assessment of surgical technical skill is central to surgical training, credentialing, and quality assurance. The Objective Structured Assessment of

⋆ Corresponding author

Technical Skill (OSATS) is a widely adopted standard in which expert raters score procedures across multiple dimensions using structured rubrics [8, 9, 2]. Despite its clinical relevance, OSATS requires manual expert review, making it labor-intensive, difficult to scale, and susceptible to inter-rater variability and subjective bias [12, 5]. Although operative videos are increasingly used for education and assessment, recent surveys highlight the lack of standardized datasets and evaluation protocols for automated surgical skill assessment [4, 6]. Importantly, expert assessment in clinical practice often includes qualitative feedback explaining assigned scores, yet this semantic information is less commonly modeled together with operative video and structured skill scores [10, 3].

Advances in computer vision have enabled analysis of surgical video, including phase recognition, tool usage, and outcome prediction [7]. Beyond surgical video analysis, generative learning has been applied to standard-plane synthesis and missing-slice imputation in cardiac MRI, joint restoration of multiple image degradations, and diffusion-based brain-tumor inpainting [17, 16, 15, 14]. However, transferring general-purpose video representations to surgical skill assessment remains challenging. Surgical training videos contain long low-information intervals, repeated movements, and subtle instrument–tissue interactions, which can dilute skill-relevant cues under uniform sampling [10, 1]. In addition, technical skill assessment is not based solely on visual appearance: expert raters often combine observed motion, tissue handling, instrument use, procedural flow, and qualitative reasoning when assigning OSATS scores. Existing automated studies have made progress in video-based assessment, but few benchmark settings explicitly connect operative video, structured multi-dimensional skill scores, and evaluator free-text comments within a unified evaluation protocol.

As an initial step toward studying these issues, we introduce SurgSkill-Bench, a pilot video-score-text benchmark for surgical skill assessment. The benchmark contains surgical training videos, six-dimensional OSATS ratings, and expert free-text comments. The primary goal of this work is to establish a reproducible benchmark setting and baseline protocol, rather than to claim a new state-of-the-art architecture. Our contributions are fourfold: (1) We introduce an initial video-score-text dataset of 214 surgical training simulation clips with six-dimensional OSATS scores and evaluator free-text comments. (2) We define two evaluation settings: video-only OSATS prediction and post hoc expert-comment-assisted OSATS prediction. (3) We provide a controlled baseline suite using representative frozen visual backbones, shared regression heads, and content-adaptive sampling. (4) We provide an explicit evaluation analysis covering thresholding, comment-assisted prediction, metadata constraints, and future participant-level validation.

2 Dataset & Benchmark

2.1 Data Collection and Composition

SurgSkill-Bench contains 214 curated surgical training simulation video clips provided by physicians. Each clip is approximately one minute long and records the

operative field during task execution. The current release is designed for post-operative surgical skill assessment rather than intraoperative decision support. Each clip is linked to six-dimensional OSATS rating events and a set of available evaluator free-text comments. Available scores are aggregated by dimension to construct video-level consensus targets. Input frames are resized to 224×224 during preprocessing. Because expert annotation of surgical skill is costly, the current release remains modest in scale. We use a video-level split to prevent direct clip overlap between training, validation, and test partitions. Specifically, 10% of videos are held out as an independent test set, and the remaining 90% are used for five-fold cross-validation. In each fold, all frames, scores, and comments derived from the same video are assigned to the same partition. Model selection is performed using the validation fold only, and the held-out test set is not used for hyperparameter selection. The reported mean and standard deviation are computed across the five video-level folds unless otherwise stated. Because participant- and session-level metadata are incomplete, the current results should be interpreted as internal video-level benchmark performance rather than participant-level generalization.

2.2 Annotations: OSATS and Expert Comments

Each video was assessed by two or more expert expert raters using six OSATS dimensions: Respect for Tissue, Time & Motion, Instrument Handling, Flow of Operation, Knowledge of Instruments, and Overall Score. Each dimension was scored on a 1–5 ordinal scale. The regression target for each dimension is the mean rater score, treated as a continuous consensus label. In addition to numerical OSATS scores, raters provided free-text comments describing technical performance, including tissue handling, instrument use, flow, motion efficiency, and overall technical quality. Since the same evaluators provided both comments and scores, comment-assisted experiments may contain score-related information and are interpreted as post hoc assisted prediction rather than autonomous video-only assessment.

2.3 Benchmark Tasks

Models are trained to regress continuous consensus scores on a 1–5 scale across the six OSATS dimensions. We report MAE and MSE as the primary regression metrics because the prediction target is the averaged rater score for each dimension. For supplementary discrimination analysis, we compute AUROC after median-based binarization for each dimension: scores greater than the median are treated as high-score samples, and scores at or below the median are treated as low-score samples. These thresholds are dataset-specific and should not be interpreted as clinically validated competency cutoffs.

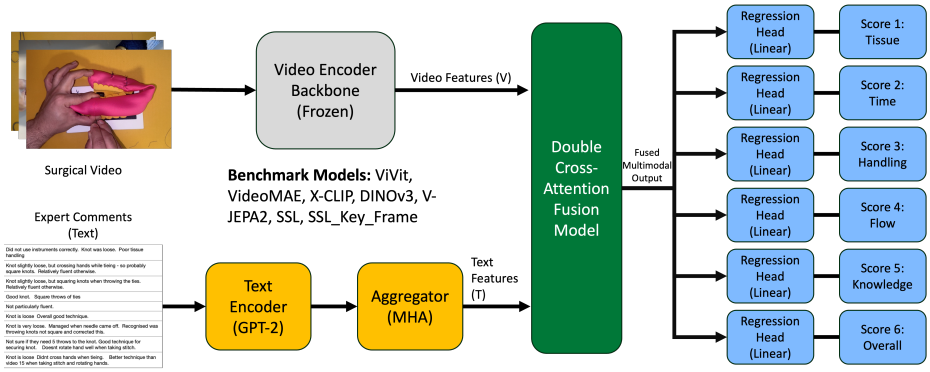


Fig. 1. Overview of the expert-comment-assisted baseline framework. Frozen video encoder backbones and a frozen GPT-2 text encoder are used to extract visual and comment-derived features. A bidirectional co-attention module fuses the two representations, followed by regression heads to predict six OSATS scores. In the video-only setting, the text branch and fusion module are removed.

3 Methodology

To evaluate baseline algorithms on SurgSkill-Bench, we provide a modular framework for video-only and expert-comment-assisted OSATS prediction. As illustrated in Fig. 1, the full assisted framework consists of three components: (1) a video representation module for extracting spatiotemporal features; (2) a text encoding module for representing evaluator comments; and (3) a fusion and regression head that combines visual and comment-derived features to predict six OSATS scores. For video-only experiments, the text branch and co-attention module are removed, and predictions are made from the video representation using the same type of regression heads.

3.1 Content-Adaptive Key-Frame Extraction

To reduce visual redundancy in surgical videos, we evaluate Content-Adaptive Key-Frame Extraction (CA-Frame) as a simple sampling baseline. A pre-trained InceptionV3 model extracts a 2048-dimensional feature vector for each frame [13]. Starting from the first frame as the reference, a new frame is selected as a key-frame when its cosine distance from the current reference exceeds a threshold $\tau = 0.05$, after which the reference is updated. If too few key-frames are detected for the required input length, the data loader falls back to uniform sampling from the original clip. Otherwise, the selected key-frame pool is sampled in chronological order to form the fixed-length input sequence required by each backbone. Standard uniform sampling and CA-Frame use the same final input length, so CA-Frame changes only the temporal locations of selected frames. In the current dataset, CA-Frame selects an average of 80 frames from over 3000

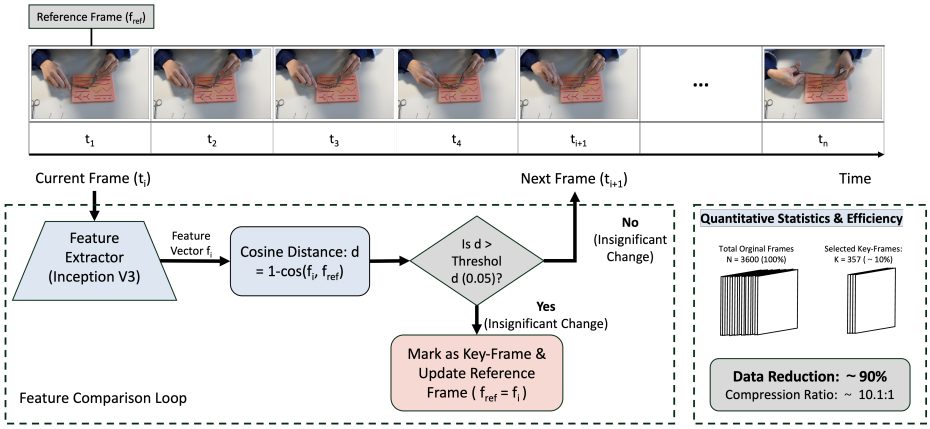


Fig. 2. Content-Adaptive Key-Frame Extraction (CA-Frame) pipeline. InceptionV3 frame features and cosine distance to a reference frame are used to select key-frames, with uniform sampling applied when key-frames are sparse.

original frames per clip. Since it uses ImageNet-pretrained InceptionV3 features, CA-Frame may miss subtle skill cues or fine-grained motion continuity.

3.2 Expert-Comment-Assisted Fusion

Expert comments are encoded using an off-the-shelf GPT-2 text encoder [11]. Token-level features are pooled into comment embeddings, and rater comments are aggregated with multi-head attention to obtain a text representation f_t. Missing comments are represented by a neutral placeholder. The GPT-2 text encoder is kept frozen in all benchmark experiments. Only the text projection layer, comment aggregation module, fusion block, and regression heads are trained.

The co-attention module is used as a controlled fusion baseline rather than a novel architectural contribution. Let $F_v \in \mathbb{R}^{T\times d}$ denote the projected video feature sequence and let $F_t \in \mathbb{R}^{M\times d}$ denote the projected comment feature sequence, where T is the number of visual tokens and M is the number of available comment embeddings for a video. The bidirectional attention is computed as:

$$A_{v\to t} = \mathrm{MHA}(F_v W_v, F_t W_t, F_t W_t), \tag{1}$$
$$A_{t\to v} = \mathrm{MHA}(F_t W_t, F_v W_v, F_v W_v). \tag{2}$$

The attended features are combined with residual connections, layer normalization, temporal pooling, and a feed-forward fusion block. In video-only experiments, the model is trained and evaluated without comments. Therefore, comment-assisted results are reported separately from video-only results.

3.3 Multi-Task Regression and Optimization

The video-only or fused representation is passed to six independent regression heads corresponding to the OSATS dimensions. The model is trained with the average MSE loss over the six consensus OSATS targets:

$$\mathcal{L}_{total} = \frac{1}{6} \sum_{k=1}^{6} ||\hat{y}_k - y_k||_2^2. \tag{3}$$

The visual backbone is frozen in benchmark runs, while the trainable projection layers, fusion module, and regression heads are optimized for six-dimensional OSATS prediction. Unless otherwise stated, both visual and textual pretrained encoders are frozen. We use regression rather than ordinal classification because the training targets are averaged rater scores, which produce continuous consensus labels. Ordinal modeling is left for future benchmark extensions.

4 Experiments

Implementation Details. All experiments are implemented in PyTorch on an NVIDIA A100 GPU. Input frames are resized to 224×224. Standard sampling uniformly samples frames from each clip, whereas CA-Frame first selects candidate key-frames and then samples them in chronological order to match the same final input length. The final input length is 16 frames for all video backbones; for DINOv3, 16 frames are encoded independently and temporally pooled. All pretrained visual and textual encoders are frozen, and only projection layers, fusion modules, and six regression heads are trained.

Models are optimized with AdamW using a learning rate of 1×10^{-4}, weight decay 1×10^{-4}, batch size 8, and gradient clipping with maximum norm 1.0. Training is performed for 200 epochs with model selection based on validation MAE. Mixed precision is used consistently across experiments. The projection and fusion hidden dimension is 256, and the co-attention module uses 4 attention heads. Results are reported across five video-level folds with fixed split seeds. The held-out test set is not used for hyperparameter selection.

Baselines. We evaluate representative pretrained visual backbones, including ViViT, VideoMAE, DINOv3, V-JEPA 2, X-CLIP, and a Surgical SSL baseline. To ensure a controlled comparison, all pretrained visual backbones are used as frozen feature extractors unless otherwise stated, and all models are evaluated with comparable projection layers and six OSATS regression heads. For image-level encoders such as DINOv3, frame-level features are temporally pooled to obtain video-level representations. For video backbones such as ViViT, VideoMAE, and V-JEPA 2, backbone outputs are projected into a shared hidden space before regression or multimodal fusion. Although X-CLIP is originally a video-language model, we evaluate it as a frozen video feature extractor under the same downstream OSATS regression protocol to ensure comparability with the other backbones. The Surgical SSL baseline uses a surgical-domain encoder pretrained on over 840 hours of publicly available surgical videos from SurgVU

Table 1. Performance of video-only models under Standard Frame Extraction and Key-Frame Extraction (Mean ± STD).

Model	Standard Frame Extraction				Key-Frame Extraction			
	MAE ↓	MSE ↓	Kappa ↑	AUROC ↑	MAE ↓	MSE ↓	Kappa ↑	AUROC ↑
VideoMAE	0.51 ± 0.03	0.41 ± 0.02	0.56 ± 0.06	0.57 ± 0.03	0.16 ± 0.01	0.04 ± 0.01	0.85 ± 0.07	0.86 ± 0.06
ViViT	0.47 ± 0.02	0.33 ± 0.02	0.61 ± 0.08	0.55 ± 0.04	0.20 ± 0.02	0.06 ± 0.01	0.81 ± 0.06	0.85 ± 0.07
X-CLIP	0.67 ± 0.11	0.64 ± 0.19	0.45 ± 0.04	0.58 ± 0.04	0.36 ± 0.07	0.26 ± 0.09	0.83 ± 0.05	0.83 ± 0.06
DINOv3	0.78 ± 0.04	0.83 ± 0.06	0.33 ± 0.06	0.61 ± 0.05	0.60 ± 0.07	0.52 ± 0.09	0.63 ± 0.06	0.79 ± 0.03
V-JEPA 2	0.48 ± 0.00	0.40 ± 0.00	0.49 ± 0.07	0.56 ± 0.06	0.15 ± 0.01	0.05 ± 0.01	0.87 ± 0.05	0.85 ± 0.04
SSL	0.72 ± 0.06	0.74 ± 0.09	0.34 ± 0.08	0.66 ± 0.05	0.34 ± 0.00	0.30 ± 0.00	0.84 ± 0.07	0.83 ± 0.04

[18].The SSL encoder was pretrained with 16-frame clips resized to 224×224 for 200 epochs using AdamW and a masked video modeling objective. No SurgSkill-Bench videos or labels were used during SSL pretraining. In downstream experiments, the SSL encoder is frozen and evaluated under the same projection-head and OSATS regression protocol as the other backbones.

Evaluation Metrics. Because OSATS scores are ordinal scores on a 1–5 scale, MAE and MSE are used as the primary regression metrics. They are computed for each OSATS dimension and then macro-averaged across the six dimensions. We additionally report quadratic weighted Cohen's kappa between rounded model predictions and rounded consensus labels as a secondary ordinal agreement statistic. For kappa calculation, continuous predictions are rounded to the nearest integer and clipped to the valid OSATS range of 1–5 before comparison with expert-derived target scores. Kappa is computed per dimension and then macro-averaged. Because the reference labels are averaged rater scores, this statistic reflects agreement with rounded consensus labels rather than agreement with individual expert raters, and should not be interpreted as human inter-rater reliability. For supplementary discrimination analysis, AUROC is computed after median-based binarization as described in Section 2. Since the median thresholds are dataset-specific and not clinically validated competency cutoffs, AUROC is interpreted as a secondary benchmark statistic. We also include a mean-score predictor as a naive baseline to contextualize MAE and MSE. This baseline predicts the training-set mean score for each OSATS dimension.

Quantitative Results. Table 1 reports video-only performance under standard frame extraction and CA-Frame extraction, and Table 2 reports the corresponding post hoc expert-comment-assisted setting. CA-Frame is associated with improved video-only performance for most backbones under the current internal video-level protocol; for example, VideoMAE improves from 0.57 to 0.86 AUROC, and ViViT improves from 0.55 to 0.85 AUROC. Evaluator comments improve several models in the post hoc assisted setting, particularly under standard frame extraction, but the benefit is not uniform across all backbones and metrics. Under key-frame inputs, ViViT and V-JEPA 2 reach the best mean AUROC of 0.88 under dataset-specific median dichotomization. These results support the assisted-prediction setting but should not be conflated with autonomous video-only assessment.

Table 2. Performance of post hoc expert-comment-assisted models under Standard Frame Extraction and Key-Frame Extraction (Mean ± STD). Expert comments are evaluator-provided feedback and may contain score-related information.

Model	Standard Frame Extraction				Key-Frame Extraction			
	MAE ↓	MSE ↓	Kappa ↑	AUROC ↑	MAE ↓	MSE ↓	Kappa ↑	AUROC ↑
VideoMAE	0.76 ± 0.01	0.80 ± 0.01	0.36 ± 0.05	0.68 ± 0.05	0.14 ± 0.02	0.03 ± 0.01	0.84 ± 0.07	0.87 ± 0.05
ViViT	0.34 ± 0.03	0.18 ± 0.03	0.83 ± 0.05	0.81 ± 0.04	0.13 ± 0.02	0.03 ± 0.01	0.85 ± 0.08	0.88 ± 0.04
X-CLIP	0.44 ± 0.03	0.29 ± 0.05	0.75 ± 0.08	0.78 ± 0.03	0.18 ± 0.01	0.06 ± 0.01	0.86 ± 0.07	0.86 ± 0.04
DINOv3	0.59 ± 0.02	0.49 ± 0.02	0.34 ± 0.06	0.75 ± 0.02	0.27 ± 0.02	0.12 ± 0.02	0.79 ± 0.09	0.84 ± 0.02
V-JEPA 2	0.56 ± 0.06	0.50 ± 0.07	0.60 ± 0.08	0.75 ± 0.03	0.13 ± 0.02	0.03 ± 0.01	0.85 ± 0.08	0.88 ± 0.04
SSL	0.73 ± 0.03	0.74 ± 0.04	0.35 ± 0.07	0.70 ± 0.02	0.16 ± 0.02	0.04 ± 0.01	0.85 ± 0.07	0.87 ± 0.03

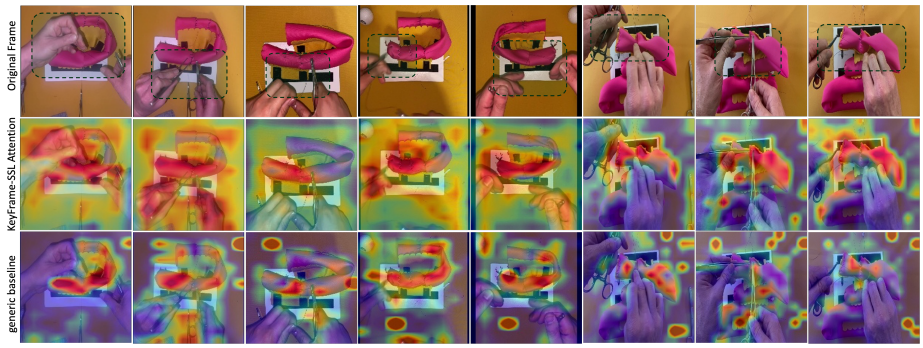


Fig. 3. Qualitative attention visualization. The first row shows original surgical frames with manual annotations of instrument, hand, and tissue interaction regions. The second row shows attention maps from the CA-Frame-guided surgical SSL encoder, and the third row shows attention maps from a generic visual baseline.

Qualitative Visualization. Fig. 3 shows self-attention heatmaps from validation samples. In these selected examples, the CA-Frame-guided SSL encoder visually appears to produce more concentrated responses around regions containing instruments, hands, and tissue interactions, whereas the generic visual baseline appears more diffuse. These visualizations are qualitative examples only and should not be interpreted as quantitative evidence of interpretability or causal proof of clinically meaningful reasoning.

5 Conclusion

We introduced SurgSkill-Bench, an initial video-score-text benchmark for surgical skill assessment that combines surgical training videos, six-dimensional OSATS scores, and evaluator free-text comments. We defined separate video-only and post hoc expert-comment-assisted settings and provided controlled baselines using frozen visual backbones, content-adaptive sampling, and video-text fusion. Under internal video-level validation, CA-Frame was associated with improved

video-only performance, while evaluator comments provided additional score-related signal in the assisted setting. These findings establish a benchmark baseline, while remaining constrained by dataset scale, incomplete participant metadata, dataset-specific AUROC thresholds, and the assisted nature of comment-based prediction. Future work will expand the dataset, add participant-level validation, include text-leakage controls, and analyze human agreement.

Disclosure of Interests

The authors have no competing interests to declare.

References

1. Avellino, I., Nozari, S., Canlorbe, G., Jansen, Y.: Surgical video summarization: Multifarious uses, summarization process and ad-hoc coordination. Proc. ACM Hum.-Comput. Interact. **5**(CSCW1) (2021). https://doi.org/10.1145/3449214, https://doi.org/10.1145/3449214
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.O.: Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine **369**(15), 1434–1442 (2013). https://doi.org/10.1056/NEJMsa1300625, https://doi.org/10.1056/NEJMsa1300625
3. De Louche, C.D., Mandal, M., Fernandes, L., Lawson, J., Bicknell, C.D., Pouncey, A.L.: Heart rate variability as a dynamic marker of surgeons' stress during vascular surgery. BJS Open **8**(5), zrae097 (2024). https://doi.org/10.1093/bjsopen/zrae097, https://doi.org/10.1093/bjsopen/zrae097
4. Dick, L., Boyle, C.P., Skipworth, R.J.E., Smink, D.S., Tallentire, V.R., Yule, S.: Automated analysis of operative video in surgical training: Scoping review. BJS Open **8**(5), zrae124 (2024). https://doi.org/10.1093/bjsopen/zrae124, https://doi.org/10.1093/bjsopen/zrae124
5. Gawad, N., Fowler, A., Mimeault, R., Raiche, I.: The inter-rater reliability of technical skills assessment and retention of rater training. Journal of Surgical Education **76**(4), 1088–1093 (2019). https://doi.org/10.1016/j.jsurg.2019.01.001, https://www.sciencedirect.com/science/article/pii/S193172041830641X
6. Lam, K., Chen, J., Wang, Z., et al.: Machine learning for technical skill assessment in surgery: A systematic review. npj Digital Medicine **5**, 24 (2022). https://doi.org/10.1038/s41746-022-00566-0, https://doi.org/10.1038/s41746-022-00566-0
7. 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 (CVPR). pp. 9522–9531 (2021). https://doi.org/10.48550/arXiv.2106.01035, https://arxiv.org/abs/2106.01035
8. Martin, J.A., 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). https://doi.org/10.1046/j.1365-2168.1997.02502.x, https://doi.org/10.1046/j.1365-2168.1997.02502.x

9. Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., Murakami, S., Saeki, S., Mukaida, H., Takiyama, W.: Using the objective structured assessment of technical skills (osats) global rating scale to evaluate the skills of surgical trainees in the operating room. Surgery Today **43**(3), 271–275 (2013). https://doi.org/10.1007/s00595-012-0313-7, https://link.springer.com/article/10.1007/s00595-012-0313-7
10. Peisl, S., Sánchez-Taltavull, D., Guillen-Ramirez, H., Tschan, F., Semmer, N.K., Hübner, M., Demartines, N., Wrann, S.G., Gutknecht, S., Weber, M., Candinas, D., Beldi, G., Keller, S.: Noise in the operating room coincides with surgical difficulty. BJS Open **8**(5), zrae098 (2024). https://doi.org/10.1093/bjsopen/zrae098, https://doi.org/10.1093/bjsopen/zrae098
11. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
12. Stulberg, J.J., Huang, R., Kreutzer, L., Ban, K., Champagne, B.J., Steele, S.R., Johnson, J.K., Holl, J.L., Greenberg, C.C., Bilimoria, K.Y.: Association between surgeon technical skills and patient outcomes. JAMA Surgery **155**(10), 960–968 (2020). https://doi.org/10.1001/jamasurg.2020.3007, https://doi.org/10.1001/jamasurg.2020.3007
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (2015). https://doi.org/10.48550/arXiv.1512.00567, https://arxiv.org/abs/1512.00567
14. Tao, T., Wang, Z., Zhang, H., Arvanitis, T.N., Zhang, L.: Diffkan-inpainting: Kan-based diffusion model for brain tumor inpainting. In: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI) (2025)
15. Zhang, L., Bronik, K., Papież, B.W.: Learning to restore multiple image degradations simultaneously. Pattern Recognition **136**, 109250 (2022)
16. Zhang, L., Pereañez, M., Bowles, C., Piechnik, S., Neubauer, S., Petersen, S., Frangi, A.: Missing slice imputation in population cmr imaging via conditional generative adversarial nets. In: (MICCAI 2019 Best Paper Finalist) International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 651–659. Springer, Cham (2019)
17. Zhang, L., Pereañez, M., Bowles, C., Piechnik, S.K., Neubauer, S., Petersen, S.E., Frangi, A.F.: Unsupervised standard plane synthesis in population cine mri via cycle-consistent adversarial networks. In: (MICCAI 2019) International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 660–668. Springer, Cham (2019)
18. Zia, A., Berniker, M., Nespolo, R., Perreault, C., Wang, Z., Mueller, B., Schmidt, R., Bhattacharyya, K., Liu, X., Jarc, A.: Surgical visual understanding (surgvu) dataset (2025). https://doi.org/10.48550/arXiv.2501.09209, https://arxiv.org/abs/2501.09209