Behavior-based Skill Assessment for Open Surgery from Multi-view and Egocentric Videos

Supplementary Material

1. Implementation Details of Baseline Models

In this section, we provide additional implementation details for the baseline models.

For the video-based baselines, we use TimeSformer-HR, which performs better than the standard TimeSformer variant. After decoding each video to 5 FPS, we construct 16-frame clips with a temporal stride of 3 frames, and sample 20 clips per video to cover the full task procedure. Frames are resized to 448×448 for TimeSformer-HR and 224×224 for VideoMAE. After extracting clip-level features, an LSTM aggregates the temporal sequence of clip features, and the hidden state is used as the final video representation. In the multi-view setting, features extracted from all camera views are concatenated before being fed into an MLP regressor for skill score prediction.

For STGCN++, as the model is applicable to longer input sequences, we uniformly sampled 500 frames from each video and use the resulting body and hand pose sequences as input. Following the original implementation, the output graph features are spatially and temporally pooled to obtain a single sequence-level feature, which is then passed to an MLP regressor. In the multi-view setting, we use two separate STGCN++ backbones to encode the body and hand pose sequences, and concatenate their output. In the egocentric setting, since body pose is unavailable, we use a single backbone to encode the 2D hand pose sequence.

2. Correlation Between Behavioral Metrics and Proficiency Rating

In Table S1, we report the absolute Spearman rank correlation coefficient between each behavioral metric in the primary exocentric framework and the ground-truth proficiency rating. These univariate correlations provide a coarse view of the association between individual metrics and skill level, but should not be interpreted as a substitute for the multivariate ablation analysis in the main paper.

Metric	f_{body}	d_h^d	d_h^n	S_h^d	S_h^n	K_h^d	K_h^n
SROCC	0.621	0.402	0.667	0.677	0.752	0.477	0.292

Table S1. Absolute Spearman rank correlation coefficients between each behavior metric and the ground-truth proficiency ratings. d and n denote dominant and non-dominant hands.

Overall, the correlation analysis suggests that multiple behavioral cues are associated with surgical proficiency, with hand smoothness exhibiting the strongest individual correlations. In particular, non-dominant hand smoothness and travel distance show relatively strong correlations, indicating that supportive hand control and motion economy may be informative for this suturing task. By contrast, the dominant-hand travel distance shows a weaker marginal association, possibly because it is more sensitive to individual execution style, such as differences in how subjects pull the suture through the graft. Body movement frequency also shows a moderate association with proficiency, consistent with the greater postural instability observed in novice subjects. The Procrustes-based consistency metrics exhibit weaker marginal correlations, especially for the non-dominant hand. However, these univariate results should be interpreted with caution, as a weaker standalone association does not preclude a metric from providing complementary information in the multivariate regression model, as reflected by the ablation analysis in the main paper.

3. Limitations

This work has several limitations. First, our dataset focuses on a single open surgical simulation task, namely suturing, and does not yet cover a broader range of procedures such as knot tying or other open surgical maneuvers. As a result, although the proposed behavioral cues are promising, their generalizability across diverse surgical tasks remains to be validated. Second, the dataset is annotated only with global skill ratings and does not provide fine-grained labels for detailed surgical events, gestures, or error types. Future work could augment the current dataset with such annotations, allowing more detailed investigations in open surgical skill assessment. Third, several proposed motion metrics rely on suturing cycle boundaries, which are currently obtained from annotated timestamps rather than estimated automatically. Finally, the current study is conducted in a simulation setting with a relatively limited number of subjects, and broader validation across larger cohorts and institutions will be an important direction for future work.

Despite these limitations, our work provides one of the first behavior-centered frameworks for open surgical skill assessment and demonstrates that explicit behavioral modeling can serve as a practical and effective alternative to heavy end-to-end video modeling, especially in small-data settings. We believe the proposed dataset, interpretable metric design, and strong empirical results establish a useful foundation for future research on holistic and data-efficient surgical skill assessment.