Supplementary Material

Daochang Liu[1,3,5], Qiyue Li [1], Tingting Jiang[1], Yizhou Wang[1,4], Rulin Miao[2], Fei Shan[2], Ziyu Li[2]
[1]NELVT, Department of Computer Science, Peking University
[2]Peking University Cancer Hospital, [3]Deepwise AI Lab
[4]Center on Frontiers of Computing Studies, Peking University
[5]Advanced Institute of Information Technology, Peking University
{daochang, liqiyue, ttjiang}@pku.edu.cn

1. Dataset Details

Detailed skill metrics are listed in Table 1. The skill metrics and the skill proxy are annotated on a Likert scale from 1 to 5. A higher score means a better skill.

Table 2 lists all the 41 event classes, which include 13 procedural events at a large granularity, 13 procedural events at a small granularity, 8 adverse events, 2 repair events, and 5 video events. The 13 large-granularity procedural events, the *"Bleeding"* from the adverse events, and the *"Camera out"* from the video events are used for skill assessment in this study.

2. Contrastive Learning Details

The temporal neighborhood $\mathcal{N}$ in the contrastive loss is set as 80 time steps with stride 10 on the simulated dataset, and 800 time steps with stride 50 on the clinical dataset.

The contrastive learning is used in the tool and event paths in our framework. The contrastive learning is disabled in the proxy path since its encoding function is set as an identity function. The contrastive learning is disabled in the visual path due to the intrinsic uncertainty in the visual features in the future. We empirically find that predicting visual features in the future could occupy too much model capacity and complicate model convergence.

Skill metrics
Gentleness
Time and Motion
Instrument Handling
Flow of Operation
Tissue Exposure
Summary Technical Skill
Summary Procedural Skill
Skill proxy
Clearness of the Operating Field

Table 1. Detailed skill annotations on our dataset.

Procedural Events (Large Granularity)	
Abdominal cavity exploration	31
Dissection of fusion tissue	19
Dissection of the greater omentum	24
LN dissection of subpyloric region (SR)	22
LN dissection of hepatoduodenal ligament region (HLR)	41
LN dissection of the superior pancreas (SP)	27
LN dissection of lesser curvature (LC)	21
LN dissection of the left gastroepiploic region (LGR)	22
Resection of the distal stomach	20
Specimen removal	20
Gastro-jejunal anastomosis	21
Jejuno-jejunal anastomosis	21
Irrigation and placement of the drains	17
Procedural Events (Small Granularity)	
Trocar insertion	22
Liver retraction	14
Peritoneal washing cytology	12
Peritoneal lesion biopsy	2
Ligation of the right gastroepiploic vein	19
Ligation of the right gastroepiploic artery	19
Ligation of the subpyloric vessel	11
Resection of the duodenum	20
Ligation of the right gastric vessel	19
Ligation of the left gastric vein	19
Ligation of the left gastric artery	20
Ligation of the posterior gastric vein	9
Ligation of the left gastroepiploic vein	18
Adverse Events	
Bleeding	279
Rupture/tear of tissue	2
Tear of spleen capsule	4
Injury/tear of serosa	8
Wrong anatomy plane	3
Perforation of the bowl	1
Adjoined organ injury	7
Insufficient pneumoperitoneum pressure	1
Repair Events	
Repair	9
Hemostasis	201
Video Events	
Camera out	352
Blurred view	128
Video exception (black screen, flip, *etc.*)	21
Software interface	12
Overlong idle time	27

Table 2. All the surgical events annotated on our dataset and the numbers of event instances.

Method ↓	SU	NP	KT	Avg.
Ours (`VTP`) 5FPS	0.791	0.761	0.784	0.779
Ours (`VTP`) 2FPS	0.715	0.760	0.792	0.757

Table 3. Impact of FPS on the simulated dataset (4-FOLD).

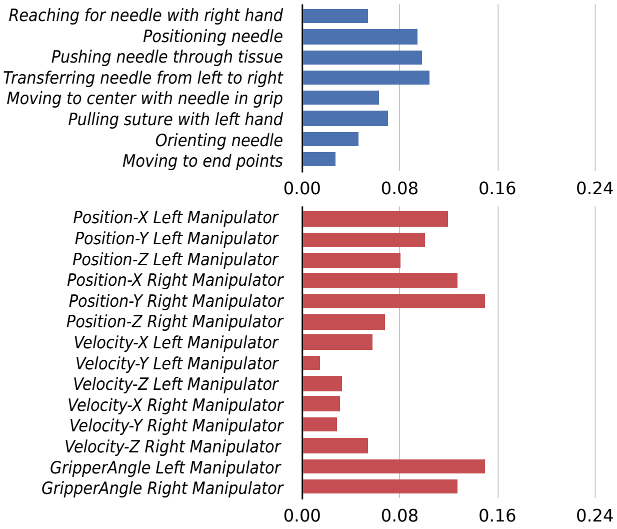

Figure 1. Blue: Correlations between model outputs and surgical gestures on the simulated needle-passing (R_{E}). Red: Correlations between model outputs and tool features on the simulated needle-passing (R_{T}).

3. Impact of FPS

The performance of our framework on the simulated dataset with the FPS reduced from 5 FPS to 2 FPS is reported in Table 3. It is shown that our framework is robust to a lower FPS.

4. R_{E} and R_{T} on the Simulated Needle-Passing and Knot-Tying Tasks

Correlations between model outputs and input features (R_{E} and R_{T}) on the simulated needle-passing task are plotted in Fig. 1. On this task, correlations are relatively lower than on the suturing task in the main paper. For R_{T}, it is observed that the position features and gripper angles have higher correlations than the velocity features.

Correlations between model outputs and input features (R_{E} and R_{T}) on the simulated knot-tying task are plotted in Fig. 2. For R_{E}, the gesture *"Reaching for needle with left hand"* has the highest correlation. For R_{T}, the position features have higher correlations than others.

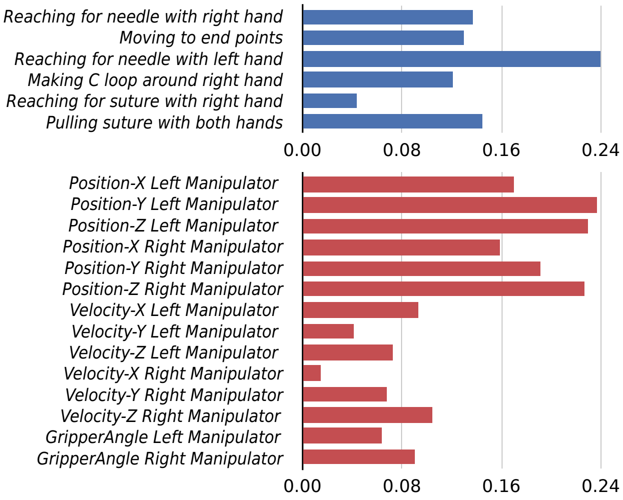

Figure 2. Blue: Correlations between model outputs and surgical gestures on the simulated knot-tying (R_{E}). Red: Correlations between model outputs and tool features on the simulated knot-tying (R_{T}).