

## Description of Additional Supplementary Files

**File Name:** Supplementary Data 1

**Description:** Heatmap of the ground-truth explanation annotations for video samples of needle handling and needle driving across hospitals.

**File Name:** Supplementary Data 2

**Description:** Precision-recall curves reflecting the alignment of different AI-based explanations with those provided by humans when assessing the skill-level of needle handling and needle driving.

**File Name:** Supplementary Data 3

**Description:** Reliability of attention-based explanations stratified across surgeon sub-cohorts when assessing the skill-level of needle handling and needle driving.

**File Name:** Supplementary Data 4

**Description:** Effect of TWIX on the reliability of AI-based explanations for the disadvantaged surgeon sub-cohort (worst-case AUPRC) when assessing the skill-level of needle handling and needle driving.

**File Name:** Supplementary Data 5

**Description:** Ablation studies and effect of TWIX when SAIS is deployed on data from a training environment.