Supplementary Material

Table 1: Multi-person tracking performance for Recording 1.

Method	IDF1 ↑	ID Switches ↓
MVPose	0.56	93
4D Association	0.58	349
TouchMap-OR (Ours)	0.47	6

Table 2: Multi-person tracking performance for Recording 2.

Method	IDF1 ↑	ID Switches ↓
MVPose	0.42	82
4D Association	0.13	137
TouchMap-OR (Ours)	0.50	81

1 Per-Recording Results

This section reports detailed quantitative results for each recording in our clinical dataset. All recordings capture the anesthesia induction phase of a surgical procedure and therefore contain largely similar activities, including preparation of monitoring equipment, patient handling, and interaction with surrounding medical devices.

The number of clinicians present varies across recordings. Recording 1 contains five people in the scene, Recording 2 contains two clinicians, and Recording 3 contains three. In addition to the primary clinicians performing the procedure, some recordings include supplementary staff performing preparatory or administrative tasks around the operating room. These additional activities introduce variations in scene complexity, including additional motion and temporary crowding around the patient area.

Tables 1–3 report multi-person tracking performance for each recording. Table 4 summarizes hand–surface contact detection and identity attribution results.

2 Implementation Details

Experiments were performed on a workstation with an NVIDIA RTX 4090 GPU, Intel i9-13900K CPU, and 64 GB RAM running Ubuntu 22.04.

We use pre-trained models for RTMO, WiLoR, and SAM2 and run all networks in inference mode without fine-tuning. RTMO is implemented using MMPose [**?**].

Table 3: Multi-person tracking performance for Recording 3.

Method	IDF1 ↑	ID Switches ↓
MVPose	0.50	61
4D Association	0.06	61
TouchMap-OR (Ours)	0.47	21

Triangulation requires at least $V_{\min} = 2$ views and uses an epipolar threshold of $\tau_{\text{epi}} = 8$ px. Joints are accepted if their mean reprojection error is below $\varepsilon_{\text{tri}} = 20$ px and their detection confidence exceeds $\tau_{\text{joint}} = 0.4$.

Depth lifting uses a patch size of $w = 5$ with a depth variance threshold $\sigma^2_{\max} = 0.4$ mm^2. Anatomical plausibility is enforced using bone-length ratios $\alpha = 0.5$ and $\beta = 2.0$.

Person tracking uses thresholds $E_{\text{init}} = 0.7$, $E_{\text{on}} = 0.8$, and $E_{\text{off}} = 0.1$, temporal decay $\lambda = 0.99$, and reuse radius $r_{\text{reuse}} = 300$ mm. Hand–person association uses a spatial gate $\tau_{\text{assoc}} = 500$ mm and motion gating with $v_{\max} = 2000$ mm/s.

Hand–surface contact detection uses EMA smoothing with $\alpha = 0.3$ and hysteresis thresholds of $\tau_{\text{on}} = 5$ cm and $\tau_{\text{off}} = 8$ cm.

Table 4: Per-recording evaluation of hand–surface contact detection and identity attribution.

Recording 1

	Episode	**Binary Contact**		**Semantic Contact**	**Identity**
Method	Recall	IoU	F1	F1	Episode ID Acc.
4D Association + heuristic	0.05	0.14	0.25	0.06	0.00
MVPose + heuristic	0.34	0.54	0.70	0.23	0.95
TouchMap-OR (Ours)	**0.62**	**0.76**	**0.86**	**0.54**	0.94

Recording 2

	Episode	**Binary Contact**		**Semantic Contact**	**Identity**
Method	Recall	IoU	F1	F1	Episode ID Acc.
4D Association + heuristic	0.00	0.05	0.09	0.00	0.00
MVPose + heuristic	0.23	0.34	0.51	0.19	0.96
TouchMap-OR (Ours)	**0.56**	**0.55**	**0.71**	**0.31**	0.95

Recording 3

	Episode	**Binary Contact**		**Semantic Contact**	**Identity**
Method	Recall	IoU	F1	F1	Episode ID Acc.
4D Association + heuristic	0.04	0.31	0.47	0.03	1.00
MVPose + heuristic	0.18	0.32	0.49	0.19	0.96
TouchMap-OR (Ours)	**0.45**	**0.55**	**0.71**	**0.26**	**0.99**