

MM-OR: A Large Multimodal Operating Room Dataset for Semantic Understanding of High-Intensity Surgical Environments

Supplementary Material

Entity	Count	Predicate	Count
Anaesthetist	14,853	Assisting	4,635
Anesthesia Eq.	4,891	Calibrating	1,721
Assistant Surg.	25,831	Cementing	48
C-Arm	731	Cleaning	113
Circulator	12,225	CloseTo	67,148
Drape	31,525	Cutting	123
Drill	2,005	Drilling	1,539
Hammer	401	Hammering	269
Head Surgeon	27,583	Holding	23,487
Instrument	17,544	Lying On	45,924
Instr. Table	32,775	Manipulating	14,273
Mako Robot	14,062	Preparing	11,681
Monitor	738	Sawing	2,383
MPS	25,895	Scanning	69
MPS Station	14,411	Suturing	132
Nurse	39,397	Touching	13,963
Op. Table	30,266		
Patient	73,671		
Saw	2,874		
Student	2,432		
Tracker	877		

Table 5. Entity and predicate counts in the scene graph annotations of the MM-OR Dataset.

Split	Timepoints	Annotations
Train	37,612	11,123
Validation	11,053	4,880
Test	13,606	2,960
Short Clips	4,725	290

Table 6. Statistics across dataset splits, timepoint and annotations.

8. Dataset Statistics

The MM-OR dataset consists of 92,983 timepoints, with 25,277 of them annotated with panoptic segmentations and scene graphs. This total is derived from 17 full-length videos, each approximately 90 minutes long, and 22 shorter clips ranging from 1 to 10 minutes each. To ensure consistency across modalities, all data streams were synchronized to a uniform rate of 1 frame per second (FPS). Table 5 provides a detailed count of annotated entities and predicates.

The dataset is divided into training, validation, test splits, and short clips summarized in Table 6.

9. Annotation Methodology

The MM-OR dataset includes over 25,000 manually annotated segmentations and scene graphs, a process that required significant effort and custom tooling. All scene graph labels were drawn from a fixed set, curated in collaboration with practicing surgeons to ensure clinical relevance. Each annotation was performed by one annotator and independently reviewed by a second annotator. On average, each scene graph comprises 8 nodes and 10 edges, with a maximum of 16 nodes and 20 edges.

10. Detailed Scene Graph Results

Predicate	Precision	Recall	F1-Score
Assisting	0.421	0.263	0.324
Calibrating	0.962	0.212	0.347
Cleaning	0.333	0.667	0.444
Close to	0.803	0.636	0.710
Cutting	0.000	0.000	0.000
Drilling	0.861	0.369	0.516
Hammering	0.708	0.548	0.618
Holding	0.792	0.409	0.539
Lying on	0.856	0.750	0.799
Manipulating	0.760	0.699	0.728
Preparing	0.699	0.845	0.765
Sawing	0.927	0.722	0.812
Scanning	0.500	0.167	0.250
Suturing	0.000	0.000	0.000
Touching	0.615	0.643	0.629
Macro Avg	0.638	0.495	0.529
Weighted Avg	0.792	0.642	0.703

Table 7. Per-predicate performance of MM2SG on the MM-OR dataset. Common predicates such as *close to* and *lying on* achieve strong performance, while rare predicates show lower scores due to limited training samples.

In Table 7, we provide more detailed results for our MM2SG model, where we report the results on each predicate. While most high to mid frequency classes can be predicted rather well the very rare predicates, such as *cutting* and *suturing* are very challenging. Our chosen macro-averaged metric emphasizes these errors by giving equal

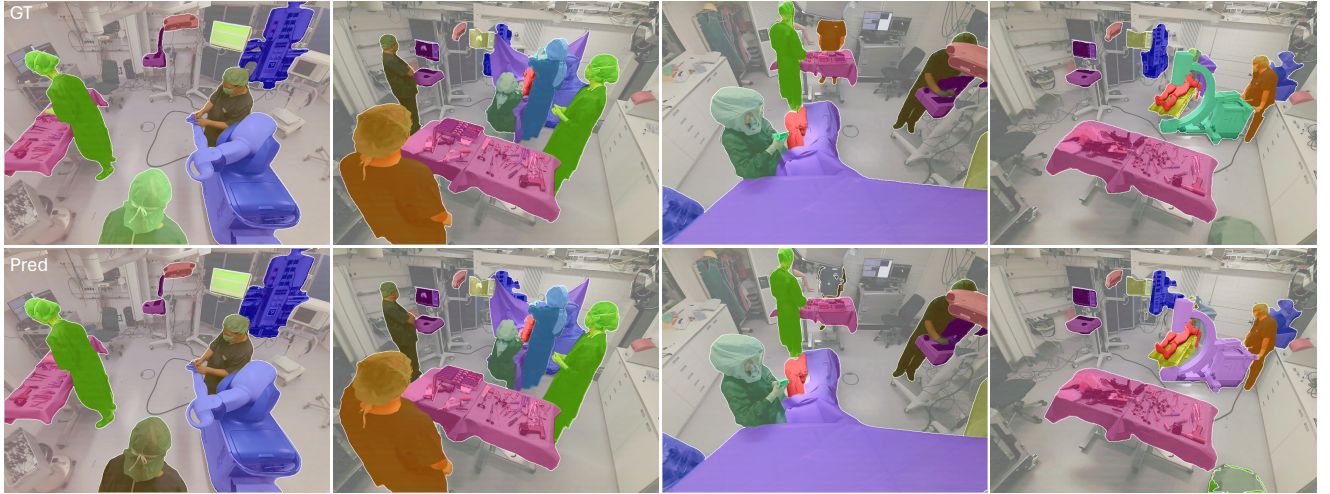


Figure 5. Qualitative segmentation results examples from a test take in MM-OR. The top row shows the ground truth segmentations and the bottom row shows the corresponding predictions.

Drop Chance	0%	25%	50%	75%
F1	0.671	0.728	0.733	0.718

Table 8. Validation results for MM2SG with varying modality drop chances. A 50% drop chance yields the highest F1-score, indicating optimal robustness.

Method	Head	Body	Tail
PSG [61]	0.473	0.156	0.052
ORacle [45]	0.690	0.456	0.120
MM2SG	0.695	0.500	0.262

Table 9. Performance of PSG, ORacle, and MM2SG across predicate frequency groups on the MM-OR dataset. MM2SG excels, particularly on rare (tail) predicates.

weight to all classes, highlighting the model’s struggles with rare predicates, whereas weighted averaging would downplay these issues by being dominated by the performance on frequent classes. We further assess the effect of varying the modality dropping chance, where modalities are randomly omitted with a given probability. Table 8 reports F1-scores on the validation set for drop chances of 0%, 25%, 50%, and 75%. The peak performance at 50% ($F1 = 0.733$) suggests that moderate modality dropping enhances generalization by reducing over-reliance on any single modality, while higher dropping (75%) slightly degrades performance. Finally, to analyze performance across predicate frequencies, we compared MM2SG against PSG [61] and ORacle [45] by grouping predicates into head ($\geq 10,000$ occurrences, e.g., *lying on*), body (1,000–10,000, e.g., *sawing*), and tail ($< 1,000$, e.g., *hammering*) cate-

gories. Table 9 shows that MM2SG consistently outperforms both baselines, with the largest relative gains on tail predicates ($F1 = 0.262$ vs. ORacle’s 0.120). This underscores MM2SG’s strength in handling rare relationships, likely due to the diverse and extensive training data in MM-OR.

11. Qualitative Segmentation Results

In Figure 5, we present qualitative examples of panoptic segmentations from multiple views. Results show that the baseline segmentation model performs quite well, however some misclassifications remain. These emphasize the potential for further improvement in complex scenarios, especially in regard to multiview and temporally consistent segmentations.

12. Overview of the Surgical Procedure

Robotic-assisted knee replacement, is a highly precise procedure designed to improve patient outcomes by optimizing implant alignment and joint function. This surgery involves replacing damaged cartilage and bone in the knee joint with artificial components to alleviate pain and restore mobility, commonly addressing conditions such as osteoarthritis. The procedure follows roughly the following workflow:

1. Preoperative CT Scan and Planning

A preoperative CT scan of the patient’s knee is used to create a 3D model of the joint. This enables the surgeon to develop a detailed surgical plan, including optimal alignment, sizing, and placement of the implants.

2. Preparation in the Operating Room (OR)

The surgical team prepares the instruments and calibrates the robotic system. The robot technician performs

operational checks and calibration, while the scrub nurse and technician drape the robot to maintain sterility. The patient is brought in, positioned supine, and the surgical site is cleaned and sterilized.

3. Tracking Array Placement and Registration

Optical tracking arrays are attached to the patient’s femur and tibia, enabling real-time tracking of their anatomy. The robotic system aligns the preoperative plan with the patient’s knee by registering key anatomical landmarks, ensuring precise guidance during surgery.

4. Bone Preparation and Implant Placement

Guided by the robotic arm, the surgeon makes precise bone cuts, assisted by haptic feedback that restricts movement to predefined boundaries. This minimizes tissue damage and ensures accurate preparation for the implants. The implants are placed according to the surgical plan, with some intraoperative adjustments.

5. Closure and Postoperative Verification

After implant placement, the surgical site is cleaned, and the wound is closed. The tracking arrays are removed, and the robotic system is shut down. To confirm proper alignment of the implants, an intraoperative X-ray scan is performed before the patient leaves the OR.

This workflow highlights the precision and integration of robotic assistance in modern knee replacement surgeries. The MM-OR dataset captures these steps in detail, providing a comprehensive resource for studying robotic-assisted surgical workflows.

13. Technical Setup

The MM-OR dataset was acquired using a multimodal recording setup designed to capture the complex dynamics of robotic knee replacement surgeries. We used multiple ceiling-mounted Azure Kinect cameras for RGB-D recordings, AXIS Q6125-LE PTZ Network Cameras for detailed views, and Sennheiser SK 300 G4-RC wireless microphone systems for the audio recordings, worn by the head surgeon, assistant surgeon and robot technician. Tracking data and robot logs were directly extracted from the robotic surgery setup, and the robot interface was recorded using an HDMI splitter. To maintain synchronization across all modalities at 1 FPS, high-resolution streams were downsampled.

14. Specimen Preparation

To ensure high realism, we used professional-grade knee phantoms from a commercial supplier². We performed realistic bone cuts and implant placements, using a new phantom for each acquisition. Fig. 6 shows the prepared phantom setup.

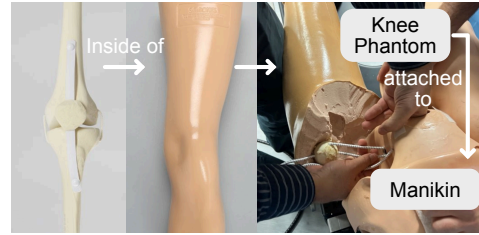


Figure 6. Prepared specimen: Bone phantom with ligaments (left), enclosed in foam for tissue simulation (middle), attached to manikin (right)

15. Supplementary Video

To provide a summary of the MM-OR dataset, we include a supplementary video. This video overlays all recorded modalities onto a dynamic 3D point cloud, offering a comprehensive visualization of the surgical environment and multimodal data. While some faces are visible in the video, no individual recognizable in the footage has any association with our institution or the authors.

²<https://www.sawbones.com>