ORacle: Large Vision-Language Models for Knowledge-Guided Holistic OR Domain Modeling

Ege Özsoy*, Chantal Pellegrini*, Matthias Keicher, and Nassir Navab

Computer Aided Medical Procedures, Technical University of Munich, Germany

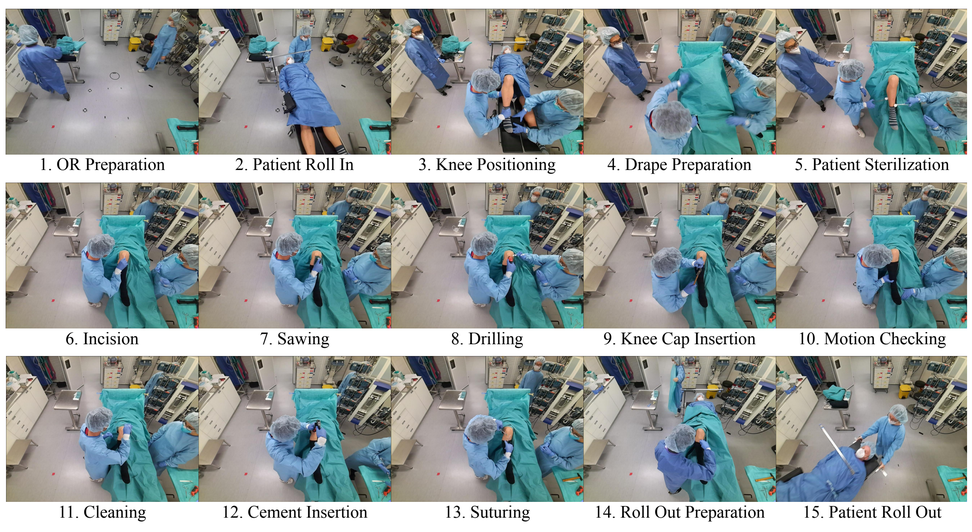
Technische Universität München

Abstract

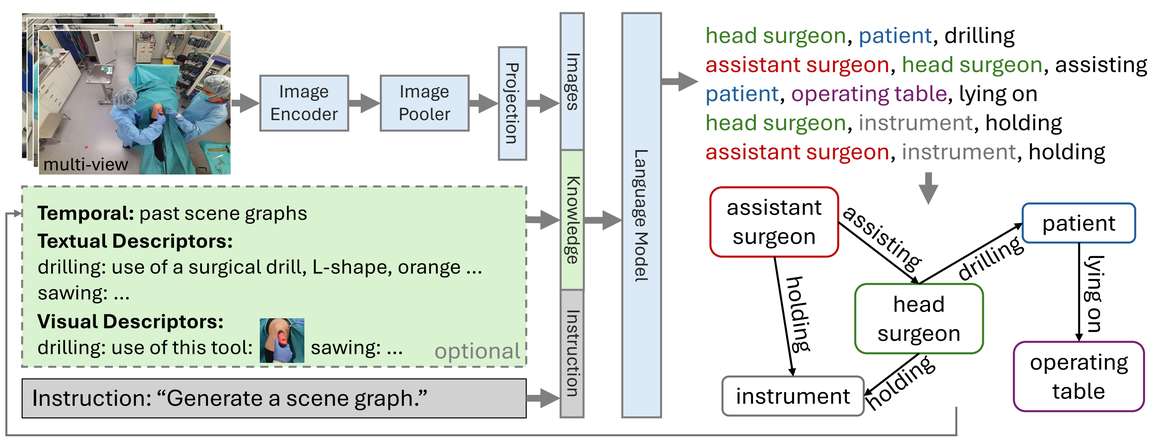
- **Context**: Every day, countless surgeries are performed worldwide, each within the distinct settings of operating rooms (ORs) that vary not only in their setups but also in the personnel, tools, and equipment used. This inherent diversity poses a substantial challenge for achieving a holistic understanding of the OR, as it requires models to generalize beyond their initial training datasets.
- **Contribution**:
  - Design the first LVLM for Holistic OR Domain Modeling
  - Formulate scene graph generation as language modeling, allowing finetuning on OR scenes
  - Design a novel multi-view image pooling module
  - Introduce a new symbolic scene graph representation
  - Propose textual and visual prompting for test-time adaptation
  - Create an automatic variability enhancement pipeline
- **Implications**:
  - Intelligently adapt to novelties in new surgeries, without retraining.
  - A pathway towards scalable, adaptable and low-cost OR Domain Modeling.

Dataset

We mainly use 4D-OR [1], which is a public OR dataset with semantic scene graph annotations for ten simulated total knee replacement surgeries. In addition, we also create an adaptability benchmark dataset, based on 4D-OR, which consists of tools, equipment and predicates not otherwise present in the dataset.



ORacle



- **Multi-View Image Pooler**: We propose a novel, transformer-based image pooler for integrating variable number of images $x_1, ..., x_N$ from multiple views, by first encoding each image with a CLIP vision encoder. The resulting embeddings are concatenated, processed by a transformer, and a joint representation is extracted from the first N tokens.
- **Scene Graph Generation as Language Modeling**: Scene graphs consist of nodes N, which correspond to entities in the scene, and edges E, which represent their relationships. We represent them as a list of triplets in the form of <*subject, object, predicate*>.
- **Symbolic Scene Graph Representation**: To handle unseen entities and predicates during inference, we represent objects and relations in a symbolic space, where subject and object are randomly assigned symbols (e.g., A, B, ...) and predicate is assigned from a set (e.g., $\alpha, \beta, ...$). For each sample, we associate these symbols with detailed descriptors in the prompt, forcing the model to rely on attributes rather than fixed class labels.
- **Textual and Visual Prompting**: Our modeling allows the integration of knowledge, either as text, by describing the following five attributes: `{object type, color, size, shape, texture}` or as images, where a single image of a respective tool or equipment can be included in the prompt, encoded using a frozen CLIP vision encoder.
- **Automatic Variability Enhancement** We generate synthetic tools from a set of attributes and place them in the OR scenes, adding variety. Despite visual imperfections, this enhances the model's ability to handle diverse descriptors.

Results

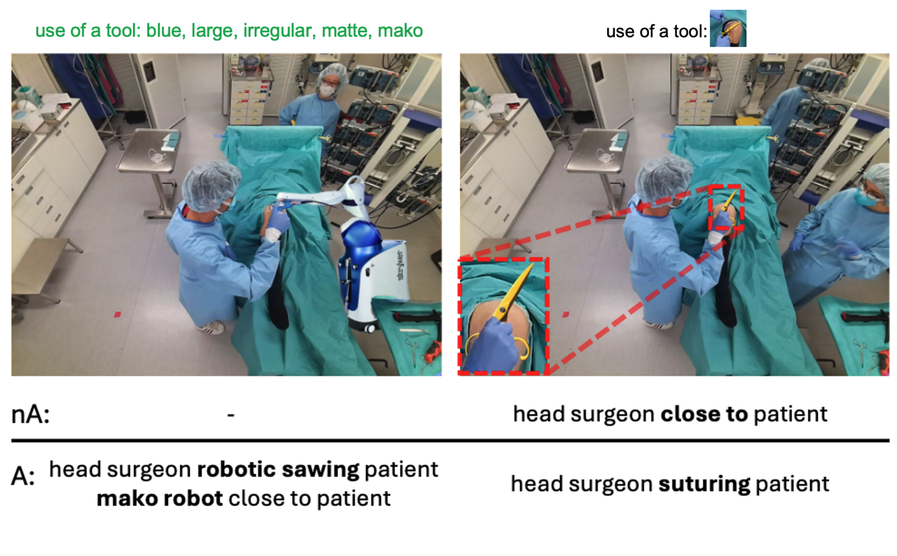
Main Results in Scene Graph Generation

Our best model achieves an F1 of 0.91, beating the existing SOTA. Additionally, we achieve this while relying on significantly less expensive sensors.

	MV	Depth	Temporal	F1
4D-OR	✓	✓	×	0.83
LABRAD-OR	✓	✓	✓	0.88
ORacle-SV	×	×	×	0.84
ORacle-SV-T	×	×	✓	0.86
ORacle-MV	✓	×	×	0.88
ORacle-MV-T	✓	×	✓	**0.91**

Adaptability Results

To evaluate adaptability, we create a benchmark of 102 images, ensuring photo-realism by considering context, orientation, and occlusions. The set includes images from two distinct views, featuring tools and equipment that vary from familiar but visually altered to entirely new types, providing a comprehensive test of the model's adaptability.



Adaptability Benchmark	Prec	Rec	F1
ORacle-MV	0.86	0.22	0.31
ORacle-adapt-Text	0.83	0.78	0.78
ORacle-adapt-Vis	0.92	0.63	0.71

Novel View Experiment	F1-2	F1-6
4D-OR	0.50	0.20
ORacle-MV-noAug	0.84	0.47
ORacle-MV	0.87	0.62

Color Switching Experiment	drill-F1	saw-F1
ORacle-MV no switching	0.94	0.98
4D-OR	0.06	0.06
ORacle-MV	0.0	0.0
ORacle-adapt-Text	0.08	0.35
ORacle-adapt-Vis	0.74	0.79

Novel Predicate Experiment	drill-F1	saw-F1
ORacle-MV full training	0.94	0.98
4D-OR	N/A	N/A
ORacle-MV	N/A	N/A
ORacle-adapt-Text	0.44	0.49
ORacle-adapt-Vis	0.79	0.52

Our extensive results show that ORacle is robust to appearance changes of known tools and equipment as well as introduction of novel entities and relations.

Conclusion

- We introduce ORacle, a novel approach leveraging large vision-language models (LVLMs) for holistic operating room (OR) domain modeling.
- Our method demonstrates both state-of-the-art results on 4D-OR and adaptability to novel tools and equipment.
- By using our symbolic representations and automatic variability enhancement pipeline, our approach can adapt to unseen entities and predicates during inference using textual or visual prompts.
- We believe this provides a pathway towards scalable, adaptable and low-cost OR Domain Modeling.

References

[1] Ege Özsoy et al. “4d-or: Semantic scene graphs for or domain modeling”. In: *MICCAI*. 2022.

[2] Ege Özsoy et al. “Labrad-or: lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms”. In: *MICCAI*. 2023.

MICCAI 2024 Marrakesh MOROCCO

Code: