

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

Ege Özsoy^{*1,2}, Chantal Pellegrini^{*1,2}, Matthias Keicher¹, Nassir Navab¹

Computer Aided Medical Procedures, Technische Universität München, Germany
Munich Center for Machine Learning (MCML), Garching, Germany

Abstract. 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. To reduce this gap, we introduce ORacle, an advanced vision-language model designed for holistic OR domain modeling, which incorporates multi-view and temporal capabilities and can leverage external knowledge during inference, enabling it to adapt to previously unseen surgical scenarios. This capability is further enhanced by our novel data augmentation framework, which significantly diversifies the training dataset, ensuring ORacle’s proficiency in applying the provided knowledge effectively. In rigorous testing, in scene graph generation, and downstream tasks on the 4D-OR dataset, ORacle not only demonstrates state-of-the-art performance but does so requiring less data than existing models. Furthermore, its adaptability is displayed through its ability to interpret unseen views, actions, and appearances of tools and equipment. This demonstrates ORacle’s potential to significantly enhance the scalability and affordability of OR domain modeling and opens a pathway for future advancements in surgical data science. Our code, pretrained models and data is publicly available at <https://github.com/egeozsoy/Oracle>.

Keywords: Semantic Scene Graph · Holistic OR Domain Modeling · Surgical Data Science · Knowledge-Guided.

1 Introduction

The Operating Room (OR) is a dynamic and complex environment, demanding precise coordination, rapid decision-making, and flawless execution for successful surgical outcomes. While there is a critical need for automatic holistic modeling of the OR to enhance surgical data science (SDS) [10,13,17], the inherent variability and complexity of surgeries present significant challenges. These range from differences in workflows and setups to the diverse tools and staff involved, complicating the development of scalable models. In the field of SDS, many

^{*} Equal contribution.

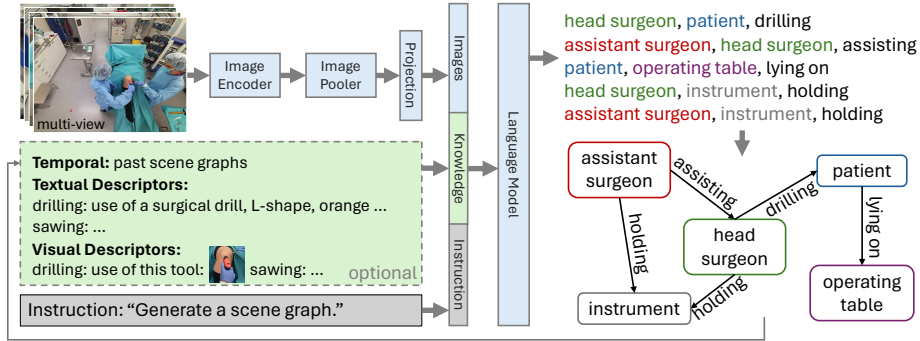


Fig. 1. An overview of our end-to-end scene graph generation architecture. ORacle takes as input multiview images and optionally additional knowledge and directly generates a scene graph token by token, considering all information at once.

works have focused on specific tasks such as detecting the surgery phase [25,4], actions [15,7,14,24], tools [9,19,27], identifying anatomies [6], or visual question answering [23,2]. Recently, Özsoy et al. [18], introduced the task of holistic OR domain modeling using semantic scene graphs, an accompanying dataset, 4D-OR, as well as a baseline method, which used the 3D point clouds and images from a timepoint to predict the scene graphs. In their follow-up work [16], the authors proposed to use memory scene graphs for efficient modeling and integration of temporality and improved the accuracy and consistency of their results. However, these approaches face limitations in real-world applicability. Their reliance on expansive hardware setups, including six 3D calibrated RGB-D cameras, poses a barrier to widespread adoption, particularly in resource-limited settings. Moreover, these methods struggle with generalization, often failing to adapt to minor changes in the OR environment, such as variations in tool color.

Simultaneously, the advent of Large Vision-Language Models (LVLMs) has revolutionized the field of computer vision by demonstrating remarkable capabilities in understanding complex visual content [1,5,12,11]. Their proficiency in processing and responding to natural language makes them ideally suited for integrating additional contextual knowledge. Yet, their potential in OR modeling remains untapped, partly due to the unique challenges of translating their capabilities to the nuanced and variable domain of surgical procedures.

To address the limitations in generalization and dependency on extensive hardware, we propose ORacle, a novel approach building upon the strengths of LVLMs for scalable and adaptable OR modeling. ORacle uniquely generates semantic scene graphs in an end-to-end manner directly from only multiview RGB images, circumventing the need for intermediate predictions or annotations such as human poses or object locations. To use multiple camera views efficiently and robustly, we introduce a multiview image pooler, which encodes a variable number of images from different perspectives into a unified representation. Furthermore, ORacle is the first method that supports the integration of

multimodal knowledge, such as temporal information and detailed descriptors of OR tools and equipment, allowing adaptation to changes not seen during training without the need for extensive re-recording and retraining. To encourage an effective use of knowledge, we design an automatic data augmentation pipeline, enhancing the variability of 4D-OR. Lastly, we create and release a challenging digitally altered subset of the publicly available 4D-OR, enabling the evaluation of the adaptability and robustness of OR modeling approaches. ORacle achieves state-of-the-art results on scene graph generation on the 4D-OR dataset and sets a strong baseline for adaptability across varied settings. By eliminating the dependency on costly depth sensors, showcasing robust performance from even a single camera perspective, and enabling knowledge guidance, ORacle paves the way for accessible, cost-effective, and adaptable holistic OR modeling, with the potential to significantly impact surgical data science.

2 Method

2.1 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. In this work, we represent them as a list of triplets in the form of $\langle \textit{subject}, \textit{object}, \textit{predicate} \rangle$ where $\textit{subject}, \textit{object} \in N$ and $\textit{predicate} \in E$. This list of triplets can be concatenated into one string, allowing us to model the scene graph generation as an image-to-sequence generation problem.

2.2 Architecture Overview

The two main components of ORacle are a visual processing module and a large language model (LLM). The visual processing module converts images from one timepoint T to a set of token embeddings in the feature space of the LLM. The LLM processes both the visual information and optionally additional knowledge, such as temporal history or textual and visual descriptors, and outputs a scene graph. A visualization of our architecture can be seen in Fig. 1.

Visual Processing We propose a novel, transformer-based [26] image pooler to enable ORacle to process a variable number of images $x_1, \dots, x_N$ from multiple views. First, each image is encoded separately using a CLIP vision encoder [22]. Then, the individual patch embeddings of all images are concatenated and the resulting sequence is processed by a transformer. Finally, we use the first N tokens as a joint representation, where N is fixed to the tokens representing one image. We then use an MLP to project this representation to the latent space of the LLM in the form of image tokens.

Large Language Model The image tokens are concatenated with the optional knowledge input and a scene graph generation instruction to form the input prompt. The LLM then autoregressively outputs a scene graph token by token.

During training, we randomly sample and shuffle the input views to ensure higher robustness to their number and order. Unlike previous methods [18,16],

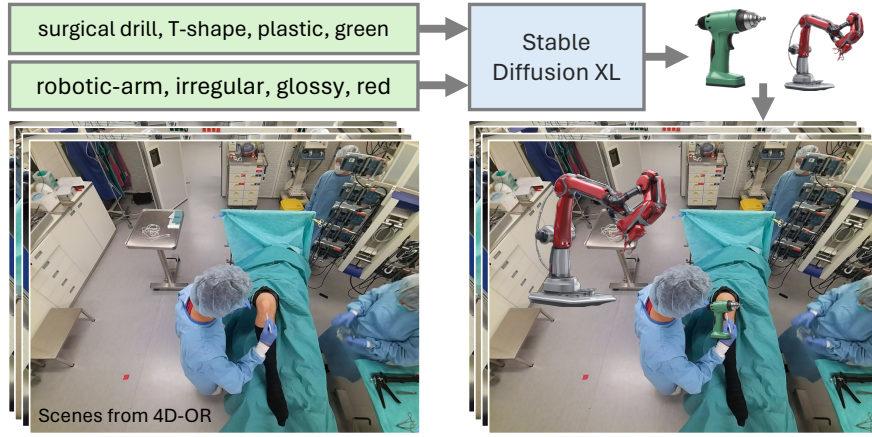


Fig. 2. An overview of our automatic variability enhancement pipeline, used during training. It first samples a set of attributes, then generates a matching object. Thereafter, it samples a suiting scene from 4D-OR and correctly places the sampled objects into it. For examples of more realistic scenes used during evaluation, see Figure 3.

ORacle is trained entirely end-to-end, and learns to directly predict a scene graph from the input, without the need for object detection or pose prediction.

2.3 Knowledge Integration

We design ORacle to allow the integration of knowledge, such as temporal context or tool and equipment descriptors, improving performance and adaptability. **Temporal Context** To efficiently integrate the surgery’s history, we condense the scene graphs from previous time points into a surgery change log, represented as a sequence of triplets. Inspired by LABRAD-OR [16], we handle the long and short-term history differently. For the long-term history, we include every triplet only once in the order they are first predicted, providing an approximate understanding of where the current time point is in the surgery. For the short-term history, we include all predictions of the last five time points, letting the model additionally know if actions are repeated and for how long they are performed. This temporal knowledge is included as text in the prompt.

Textual Descriptors The goal of textual descriptors is to inform the network of appearances of tools and equipment by describing their attributes in text form. We define an attribute space consisting of the following five attributes: {**object type**, **color**, **size**, **shape**, **texture**}. In this space, we can flexibly describe relations and entities. Such descriptors allow to inform the network about new or different-looking objects during inference time without the need for retraining.

Visual Descriptors With textual descriptors, it can be difficult to accurately describe appearances. Therefore, we additionally introduce the concept of visual descriptors, where we include a single image of the respective tool or equipment

Symbolic Scene Graph Representation We want to enable our model to work with unseen entities and predicates during inference in an open vocabulary setting and pay more attention to the descriptors. To accomplish this, in our adaptable models, we represent all objects and relations in a symbolic space. For instance instead of $\langle head\ surgeon, patient, drilling \rangle$, we represent a triplet as $\langle SUB, OBJ, PRED \rangle$ with $SUB, OBJ \in \{A, B, \dots, Z\}$ and $PRED \in \{\alpha, \beta, \dots, \omega\}$. For every sample, we randomly pick symbols for all objects and predicates and pair every symbol with its descriptor in the prompt. This forces the model to rely on the descriptors instead of class names to make a correct prediction. At inference time, we can guide the model to understand appearance changes and recognize novel objects or predicates by adapting the descriptors.

3 Experiments

Dataset We train and evaluate on 4D-OR [18] with the official data splits. 4D-OR includes ten takes of simulated partial knee replacement surgeries, recorded from 6 views with RGB-D cameras. ORacle only uses four views without depth. All 6734 scenes are labeled with semantic scene graphs, phases, and human roles.

Table 1. Distribution of included entities and relations in our adaptability benchmark. Relations and entities not existing in 4D-OR are marked with *.

	cementing	cutting	drilling	hammering	sawing	suturing	Kuka*	Mako*	rob. sawing*
Nr.	12	12	12	12	12	12	10	10	10

Table 2. Scene graph generation results on the 4D-OR test set. SV and MV indicate single view or multiview input, respectively.

	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

Table 3. Ablation study on adaptable models, showing using symbolic representation and adaptable knowledge representations do not lead to worse results on the 4D-OR test set.

	F1
ORacle-MV	0.88
ORacle-adapt-Text	0.87
ORacle-adapt-Vis	0.88

photo-realism by accounting for the appropriate context, orientation, and occlusions, distinguishing it from the outputs of our automatic pipeline. It comprises images captured from two distinct views and features a spectrum of tools and equipment ranging from familiar yet visually altered to entirely novel types. Each sample consists of a positive image with the digital alteration and a negative image without the alteration, as well as a prompt textually and visually describing the adaptations. The included relations and entities can be found in Table 1.

Evaluation Metrics Following previous works, for scene graph prediction, we evaluate our model using macro F1 over all relation types and temporal consistency between the time points. Additionally, we report precision and recall for evaluation on our adaptability benchmark.

Implementation Details We initialize our LVLM with the weights of LLaVA-7B [12], with a CLIP [22] based image encoder and Vicuna-7B [3] as the LLM. We fully finetune the image feature extractor and finetune the LLM using LoRA [8] on a Nvidia A-40 GPU. We train all our models until convergence, either on the original 4D-OR dataset or the samples generated by our automatic variability enhancement pipeline. For the temporal models, we closely follow LABRAD-OR [16], including using curriculum learning, whereby a model previously trained without temporal information is further finetuned using temporal knowledge.

Table 4. Results on our 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

Table 5. Results when only using view two (F1-2) or six (F1-6). "noAug" indicates training without order augmentation.

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

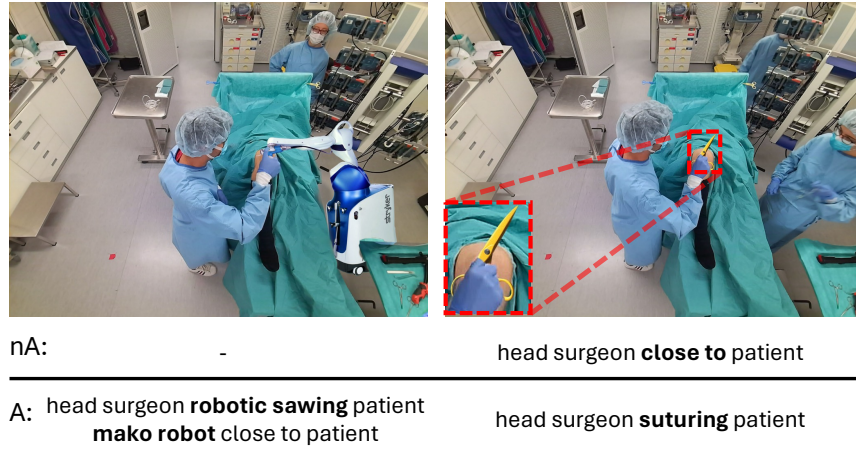


Fig. 3. Results on our adaptability benchmark of the non-adaptable ORacle-MV model (nA) and our adaptable models (A). Left: ORacle-adapt-text; Right: ORacle-adapt-vis.

4 Results and Discussion

Scene Graph Generation We compare ORacle to previous works on 4D-OR in Table 2. Our best-performing model, relying on four RGB images and temporal context, outperforms the previous SOTA, which additionally relies on two more RGB images and six depth images, by 3%. ORacles’ predictions also are temporally consistent, reaching a consistency score of 0.89 compared to 0.87 in LABRAD-OR [16], getting very close to the ground truth consistency of 0.9. Without temporality, we outperform 4D-OR [18] by 5%, and get comparable results even when relying on one single view. Overall, the results show that both temporality and multi-view integration positively affect performance. In Table 3, we show the results of our adaptable model variants, demonstrating that performance remains similar when using the symbolic models with descriptors.

Downstream Tasks We evaluate our model on clinical role prediction and surgical phase recognition. Unlike prior works, ORacle always outputs a scene graph already including clinical roles, alleviating the need to explicitly solve clinical role prediction. It achieves a macro F1 of 0.85, which is only slightly worse than 0.89 from LABRAD-OR [16], which relies on ad-hoc analyses, including tracking and a secondary role prediction model. We also compare our model to the existing SOTA in surgical phase recognition, closely following their methodology [17]. Without adapting any of the heuristics for mapping scene graph sequences to surgical phases, we achieve an F1 score of 0.99, compared to their result of 0.97.

Adaptability We validate the robustness and adaptability of our method when using textual and visual descriptors, depicted as ORacle-adapt-text and ORacle-adapt-vis, respectively. We work in a single time point setting, therefore comparing the non-temporal ORacle models to the 4D-OR method [18].

Table 6. Results when the colors of the drill (orange) and saw (green) were digitally altered to be the exact opposite of each other only during evaluation. "no switching" is provided as a baseline value.

	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

Table 7. Results when training without drilling or sawing scenes. "full training" corresponds to using these scenes for training. Both 4D-OR and ORacle-MV can not predict any unseen predicates, indicated by N/A.

	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

Single View: Table 5 analyzes the robustness to using a single view during inference. We evaluate all models on only view two, which was included in the training views, and see that ORacle shows high robustness to this single-view setting, reaching an F1 score of 0.87, which is only 1% less than using all four views. Secondly, we compare performance on view six, which is a novel view for ORacle, not used during training. Even though 4D-OR did use this view in training, ORacle performs significantly better, demonstrating its high robustness to novel views. Further, we ablate the effect of order augmentations during multiview training, confirming that these are crucial for increasing robustness.

Novel appearances and objects: Table 4 shows the results on our adaptability benchmark, and Figure 3 shows qualitative examples. Both adaptable models perform significantly better than the non-adaptable model variant. This shows both robustness to appearance changes of known tools and equipment as well as to novel ones. We also digitally altered one take of 4D-OR, switching the color of the drill and the saw in all frames. While both 4D-OR and ORacle-MV get confused in this setting, our adaptable models, especially when using visual descriptors, are substantially more robust towards this change, as shown in Table 6. Lastly, we test the understanding of novel relations by training ORacle on a modified training set, excluding all time points in which the drilling or sawing relation occurs. Prior works and our non-adaptable model, cannot predict relations that were not seen during training. Our adaptable models show a good performance on these relations as shown in Table 7.

5 Conclusion

In this work, we propose ORacle, an adaptable knowledge-guided holistic OR modeling approach building upon Large Vision-Language Models. Our method generates semantic scene graphs end-to-end from multiview RGB images, by employing a novel multiview image pooler. We introduce multimodal knowledge guidance, which for the first time, allows adapting to previously unseen concepts

in inference time. Our results show that ORacle reaches state-of-the-art results not only in the benchmark task of scene graph generation but also demonstrates significantly improved adaptability to unseen OR scenarios. By reducing the requirement for extensive data collection and annotations, we believe ORacle opens up a pathway for more affordable and scalable OR domain modeling.

6 Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

7 Acknowledgements

This work has been supported by Software Campus & Bundesministerium für Bildung und Forschung (BMBF) with grant [ZN 01IS17049]. The authors have been also partially supported by Stryker, the EVUK programme ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022)
2. Bai, L., Islam, M., Ren, H.: Revisiting distillation for continual learning on visual question localized-answering in robotic surgery. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 68–78. Springer (2023)
3. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023)
4. Czempel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Shenzhen, China, October 4-8, 2020, Proceedings, Part III. Lecture Notes in Computer Science*, vol. 12263, pp. 343–352. Springer (2020)
5. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
6. Das, A., Khan, D.Z., Williams, S.C., Hanrahan, J.G., Borg, A., Dorward, N.L., Bano, S., Marcus, H.J., Stoyanov, D.: A multi-task network for anatomy identification in endoscopic pituitary surgery. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 472–482. Springer (2023)
7. Hao, L., Hu, Y., Lin, W., Wang, Q., Li, H., Fu, H., Duan, J., Liu, J.: Act-net: Anchor-context action detection in surgery videos. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 196–206. Springer (2023)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021)
9. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., et al.: Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II* 27. pp. 218–229. Springer (2021)
10. Lalys, F., Jannin, P.: Surgical process modelling: a review. *International Journal of Computer Assisted Radiology and Surgery*, Springer Verlag **9**, 495–511 (2014)
11. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023)
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023)
13. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kennigott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G.D., Jannin, P.: Surgical data science for next-generation interventions. *Nature Biomedical Engineering* **1**(9), 691–696 (Sep 2017)

14. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N.: Encoding surgical videos as latent spatiotemporal graphs for object and anatomy-driven reasoning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 647–657. Springer (2023)
15. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* **78** (2022)
16. Özsoy, E., Czempiel, T., Holm, F., Pellegrini, C., Navab, N.: Labrad-or: Lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2023)
17. Özsoy, E., Czempiel, T., Örnek, E.P., Eck, U., Tombari, F., Navab, N.: Holistic or domain modeling: a semantic scene graph approach. *International Journal of Computer Assisted Radiology and Surgery* (2023). <https://doi.org/10.1007/s11548-023-03022-w>, <https://doi.org/10.1007/s11548-023-03022-w>
18. Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII. Springer (2022)
19. Paranjape, J.N., Sikder, S., Patel, V.M., Vedula, S.S.: Cross-dataset adaptation for instrument classification in cataract surgery videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 739–748. Springer (2023)
20. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
21. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Gool, L.V.: Highly accurate dichotomous image segmentation. In: ECCV (2022)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
23. Seenivasan, L., Islam, M., Kannan, G., Ren, H.: Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 281–290. Springer (2023)
24. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 505–514. Springer (2023)
25. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**(1), 86–97 (2016)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
27. Wang, R., Ktistakis, S., Zhang, S., Meboldt, M., Lohmeyer, Q.: Pov-surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 440–450. Springer (2023)