

CAT-SG: A Large Dynamic Scene Graph Dataset for Fine-Grained Understanding of Cataract Surgery

Felix Holm^{1,2,3}[0000–0002–8493–9941], Gözde Ünver^{1,2}[0009–0000–5219–8926], Ghazal Ghazaei²[0000–0002–0792–3583], and Nassir Navab^{1,3}

¹ Chair for Computer-Aided Medical Procedures, Technical University Munich, Germany

² Carl Zeiss AG, Munich, Germany

³ Munich Center for Machine Learning (MCML)
`felix.holm@tum.de`

Abstract. Understanding the intricate workflows of cataract surgery requires modeling complex interactions between surgical tools, anatomical structures, and procedural techniques. Existing datasets primarily address isolated aspects of surgical analysis, such as tool detection or phase segmentation, but lack comprehensive representations that capture the semantic relationships between entities over time. This paper introduces the Cataract Surgery Scene Graph (CAT-SG) dataset, the first to provide structured annotations of tool-tissue interactions, procedural variations, and temporal dependencies. By incorporating detailed semantic relations, CAT-SG offers a holistic view of surgical workflows, enabling more accurate recognition of surgical phases and techniques. Additionally, we present a novel scene graph generation model, CatSGG, which outperforms current methods in generating structured surgical representations. The CAT-SG dataset is designed to enhance AI-driven surgical training, real-time decision support, and workflow analysis, paving the way for more intelligent, context-aware systems in clinical practice. The dataset is available at github.com/felixholm/CAT-SG.

Keywords: Scene Graphs · Surgical Data Science · Dataset · Cataract Surgery

1 Introduction

Advancements in machine learning and computer vision are transforming computer-assisted surgery, enabling precise planning, real-time decision support, and automated workflow analysis. While progress has been made in surgical video analysis for tasks like tool detection, phase segmentation, and triplet recognition, achieving a holistic yet fine-grained understanding of surgical workflows remains a challenge.

Scene graphs offer a robust framework for modeling structured interactions within complex environments, representing entities as nodes and their relationships as edges. In computer vision research, they are increasingly recognized as

valuable tools for providing structured, fine-grained, and human-readable representations that yield holistic insights into dynamic scenes [11,14,16]. Therefore, the application of dynamic scene graph representations holds significant promise for revolutionizing surgical workflow analysis by encoding surgical scenes through surgical elements, their spatio-temporal features, and interactions. Although prior research has explored scene graph-based approaches for surgical workflow recognition [15,10,12], these efforts have been constrained by the limitations of existing datasets lacking fine-grained semantic interactions necessary for detailed procedural analysis.

Cataract surgery, one of the most common and intricate ophthalmic procedures, exemplifies the need for structured surgical representations. The success of phacoemulsification, a widely used cataract removal technique, depends on precise tool-tissue interactions, such as controlled ultrasonic energy application for lens fragmentation and delicate intraocular manipulation. Cataract procedures involve fine-grained motion dynamics and subtle anatomical interactions, which existing datasets fail to systematically capture, hindering the development of automated workflow analysis solutions.

To address this gap, we introduce the Cataract Surgery Scene Graph (CAT-SG) dataset, a novel resource derived from the CATARACTS dataset [1]. CAT-SG provides structured representations of surgical elements, encompassing their geometric, temporal, and semantic interactions, with a particular focus on tool-tissue relationships. It also introduces a new downstream task of surgical technique recognition, complementing the existing surgical phase annotations in CATARACTS.

Our contributions are summarized as follows:

CAT-SG Dataset: We present the first Cataract Surgery Scene Graph dataset, with over 1.8 Million annotated relations. Additionally, CAT-SG provides novel annotations for surgical techniques, enabling a more comprehensive understanding of procedural variations beyond traditional phase recognition.

Comprehensive Benchmarking and Evaluation: To demonstrate the utility of CAT-SG, we establish benchmarks for the three key tasks of our dataset: scene graph generation, phase recognition and technique recognition, demonstrating the dataset’s applicability across various surgical workflow tasks.

Novel Scene Graph Generation Model (CatSGG): Additionally, we propose CatSGG, a new scene graph generation model leveraging large-scale pre-training and spatio-temporal attention to efficiently generate structured representations of surgical interactions outperforming the state-of-the-art method ORacle [23] on CAT-SG.

2 CAT-SG Dataset

2.1 CATARACTS

We introduce a novel dataset named CAT-SG that extends the publicly available CATARACTS dataset [1] with scene graph annotations. The CATARACTS

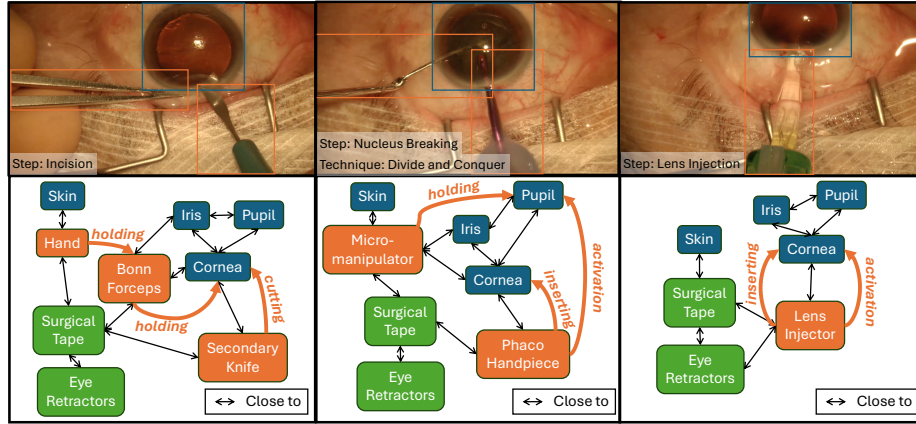


Fig. 1. Samples from one video of CAT-SG, showing Image and Scene Graph including semantic and geometric relations

dataset consists of 50 high-resolution videos of cataract surgeries, collected at Brest University Hospital. Each video captures a complete surgical procedure with a duration of 11 minutes on average (min: 6 min, max: 40 min). With over nine hours of annotated surgical footage, CATARACTS provides a robust basis for developing machine learning models for surgical video understanding [1].

2.2 Annotations

The scene graph annotations were created through a manual labeling process conducted by 9 trained student annotators. The annotators reviewed the videos and interactively marked each instance of a surgical tool interacting with an anatomical structure, followed by an iterative expert review process to verify the validity of our labels. After over 1200 combined hours of annotating, this process resulted in a detailed temporal mapping of tool-tissue interactions throughout the surgeries. The interactions are classified into 8 categories shown in Table 1.

The following 29 objects were considered for annotations: Pupil, Surgical Tape, Hand, Eye Retractors, Iris, Skin, Cornea, Hydrodissection Cannula, Vis-

Table 1. Relations annotated in CAT-SG

Relation	Example	Samples
Holding	e.g., forceps fixating sclera	13,380
Activation	e.g., phaco-handpiece using emulsification	44,552
Pushing	e.g., micro manipulator pushing & rotating nucleus	3,874
Pulling	e.g., cystotome pulling on capsular bag flap	11,895
Cutting	e.g., corneal incisions using knife	1,925
Inserting	tool is inserted or retracted through the primary or	34,016
Retracting	secondary incision in the cornea	23,886
Close to	spatial proximity of one object to another	1,677,724

Table 2. Size of comparable datasets and CAT-SG

Dataset	Videos	Avg. Duration	Annotated Frames	Unique Objects	Unique Relations	Annotated Relations
4D-OR	10	11 min	6,743	11	14	103,740
CholecT45	45	33 min	90,489	20	9	127,385
CAT-SG	50	11 min	164,162	29	9	1,811,252

coelastic Cannula, Capsulorhexis Cystotome, Rycroft Cannula, Bonn Forceps, Primary Knife, Phacoemulsification Handpiece, Lens Injector, Irrigation/Aspiration Handpiece, Secondary Knife, Micromanipulator, Capsulorhexis Forceps, Suture Needle, Needle Holder, Charleux Cannula, Vitrectomy Handpiece, Mendez Ring, Marker, Troutman Forceps, Cotton, Iris Hooks, Vannas Scissors.

We sample and annotate the 50 videos in our dataset at a temporal resolution of 5 fps, making CAT-SG the largest dataset among comparable efforts for detailed surgical scene understanding (Table 2). We also use pseudo-labelled segmentation masks based on the CaDIS dataset to provide a grounding to each object in our annotations (position, size, bounding box). We will also release these masks in addition to our dataset.

Use Cases: CAT-SG is designed to support detailed surgical video understanding through structured scene graph representations offering the following tasks:

Surgical Scene Graph Generation: This task involves automatically constructing structured scene graphs from surgical videos. Given a sequence of frames, the goal is to detect surgical instruments, anatomical structures, and their interactions over time. Successful models must not only recognize objects but also infer their relationships, enabling a deeper understanding of surgical workflow and tool usage patterns.

Surgical Workflow Recognition: Our dataset enables surgical workflow recognition by leveraging the predefined 19 surgical steps from the CATARACTS dataset [1]. With the integration of scene graph annotations, this process gains additional structure and granularity, reinforcing its role as a fundamental task in surgical video understanding.

Surgical Technique Recognition: To enhance surgical video analysis, we add new annotations distinguishing two nucleus-breaking techniques: “Stop and Chop” and “Divide and Conquer.” Each procedure in the dataset is labeled as using one of these two techniques during the nucleus breaking step. Identifying these techniques requires a nuanced understanding of tool movements and procedural dynamics, making this a complex and valuable benchmark for automated surgical workflow analysis.

3 CatSGG Scene Graph Generation Method

We introduce a novel scene graph generation method comprising of two steps: 1) entity localization and base graph construction, 2) geometric and semantic

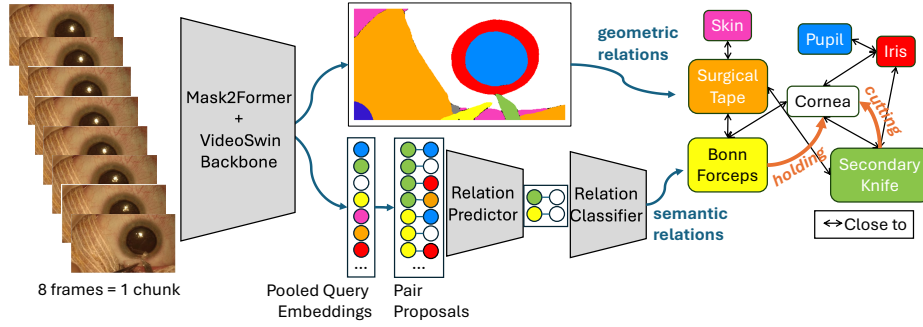


Fig. 2. Overview of the CatSGG+ Pipeline: an input video chunk is fed into Mask2Former with a video-based backbone to extract rich spatio-temporal features. The extracted query embeddings are pooled and used to generate pair proposals, which are then processed by a relation predictor and classifier to infer semantic relations. Geometric relations are inferred from the predicted mask.

relation prediction.

Annotation-efficient Entity Localization While bounding boxes provide coarse localization, they lack the spatial precision needed for modeling surgical interactions in scene graphs. We leverage segmentation models instead, as they offer fine-grained instance localization, enabling more accurate spatial reasoning and relationship extraction. Surgical video segmentation could also be challenging due to the long duration of procedures, e.g. cataract surgeries typically last around 11 minutes, making manual annotation and computational processing at high fps highly burdensome. To address this, we propose an annotation-efficient solution for instance segmentation of surgical elements within CATARACTS frames by leveraging state-of-the-art segmentation and video-language pretraining. Our approach builds upon Mask2Former (M2F) [4] while incorporating domain-specific prior knowledge from Watch&Learn [8], a large-scale video-language pretraining framework. Specifically, we replace the standard backbone of M2F with a finetuned VideoSwin [13], which was trained within the VALOR framework [3] alongside a BERT encoder [6] for video-language alignment. We opt for this model over the original VideoSwin because its pretraining on 2,900 YouTube cataract surgery videos [8] provides strong domain priors, allowing effective segmentation with fewer annotated samples. Moreover, the spatio-temporal design of VideoSwin enhances the model’s awareness of cataract surgery dynamics, further improving segmentation performance. By integrating VideoSwin into M2F, we adapt the segmentation pipeline to better capture the temporal and contextual dependencies in surgical videos while requiring fewer annotated frames.

Geometric Relation Prediction: The “Close to” relation, which captures spatial proximity, is determined based on M2F segmentation predictions. Specifically, we identify instances with adjacent masks (touching boundaries) as being

close to each other, following the approach in [10].

Semantic Relation Prediction: The VideoSwin backbone of our M2F processes 8 consecutive frames as a chunk, with segmentation and scene graph generation performed only on the last frame. Consequently, our backbone inherently captures temporal information across frames. Inspired by query embedding solutions [17,18,20,22,21,19], we leverage the output query vectors (q) from M2F for relation prediction. Each query vector encodes class-specific and spatial information for a single instance. To model pairwise relationships, we construct pair proposal embeddings by concatenating the queries of two instances. Since semantic relations occur only between tool-tool or tool-anatomy pairs, we incorporate dataset priors to construct these pair proposals efficiently. Relation prediction consists of two tasks: (1) detecting whether a relation exists ($e \geq 0.5$), and (2) classifying the relation type. Following [20], we employ: A 2-layer model for binary relation existence prediction ($f_{\text{existence}}$), followed by a 3-layer multi-label classification model ($f_{\text{classification}}$) for relation type prediction.

$$\begin{aligned} \text{pair}_n &= [q_i; q_j] \text{ where } i \neq j, i \in N_{\text{tool}}, j \in N_{\text{tool}} \cup N_{\text{anatomy}} \\ e &= \sigma(f_{\text{existence}}(\text{pair}_n)), e \in [0, 1] \\ \text{if } e &\geq 0.5, c = f_{\text{classification}}(\text{pair}_n) \end{aligned} \quad (1)$$

We use binary cross-entropy loss with sigmoid activations for both models.

Both models operate on the pair embeddings derived from the query vectors of the last frame in the chunk. We refer to this method as CatSGG. To further enhance temporal consistency, we extend this approach by aggregating query vectors from all frames in the chunk. Inspired from [20], we apply max pooling on same-class queries across these frames to obtain a single representative vector per class.

These temporally enriched vectors are then used to create pair embeddings for relation predictions. We refer to this temporally aware extension as CatSGG+.

4 Experiments

CatSGG: Semantic Segmentation: From preliminary experiments we find that training M2F with the pretrained VideoSwin backbone from [8] achieves equal performance with much faster convergence and improved generalizability by subsampling the training data to 18 chunks per training video. Using this configuration we train M2F on the 29 classes from CAT-SG, by merging pseudo-masks based on CaDIS [9] with CAT-SG annotations. The model achieves a mIoU of 92.12%, demonstrating its ability to generate initial scene graph nodes.

CatSGG: Semantic Relation Prediction: Using the frozen pretrained segmentation model, we train the relationship prediction models with 18 chunks per training video. Since semantic relations are relatively infrequent, we apply sampling constraints to ensure that each selected chunk contains at least one semantic relation. We apply a 0.5 threshold to filter predictions from both the

relation existence and classification models.

ORacle [23]: We use ORacle as a benchmark, representing the most recent SOTA method in scene graph generation within the surgical setting. ORacle is a large vision-language model and does not require bounding boxes or scene graph grounding, which makes it highly versatile and easily adaptable to various domains. ORacle is originally developed for the 4D-OR dataset, consisting of external views captured by overhead cameras in the operating room (OR). We train and evaluate ORacle’s single view variant, both with and without temporality (denoted as ORacleSV and ORacleSVT, respectively). To account for the distribution of relation classes, we report F1 scores for each class, as well as micro- and macro-averaged F1.

Downstream Task Baselines: Since semantic scene graphs connect the graph and language domain, they can be used both in Graph Neural Networks (GNNs) and Large Language Models (LLMs). We present baselines for both approaches, leveraging our scene graphs to recognize surgical workflows and techniques. We report Accuracy and F1 score as the well-established metrics for these tasks [7]. **GNN:** We use a 3-layer GATv2 [2] model as our GNN baseline. Like [10], we encode node groundings (i.e., position and size in the image) as part of the node features. Temporality is captured by connecting scene graphs from different timesteps with temporal edges between nodes of the same class, creating a dynamic scene graph.

LLM: We fine-tune a Llama 3.2 3B model for our downstream tasks using QLoRA and its recommended parameters [5]. To represent our scene graphs, we prompt the model with a list of entities and relations, along with their groundings provided as text (e.g., “<Object> at (<x>, <y>) with size <size>”). Temporality is incorporated through a history mechanism: the prompts include scene graphs from the last N steps prior to the current scene graph that needs to be classified.

5 Results & Discussion

Scene Graph Generation. Table 3 presents the results for scene graph generation on CAT-SG. Our CatSGG outperforms the SOTA ORacle by over 8 percentage points (p.p.) on Macro-F1. The inclusion of longer-term temporal context in CatSGG+ further improves performance, allowing the model to better capture the dynamically evolving semantic relations over time. Surprisingly, adding temporality to ORacle (ORacleSVT) leads to a performance drop. We hypothesize this is due to ORacle aggregating its predictions over time, which can propagate mispredictions, thereby degrading overall performance. A video of qualitative results for CatSGG+ is attached in the supplementary.

Surgical Workflow Recognition. As shown in Table 4, introducing semantic relations leads to improvements in both accuracy and F1, surpassing the baseline from [10] by over 5 p.p. This demonstrates that explicitly encoding

Table 3. Scene Graph Generation Results on CAT-SG.

	F1 Score for each Relation									Overall	
	Close to	Holding	Activation	Pushing	Pulling	Cutting	Inserting	Retracting	"none"	Micro F1	Macro F1
ORacleSV	67.03	3.74	41.76	1.57	33.08	31.06	39.75	15.00	78.83	72.45	34.65
ORacleSVT	57.64	19.82	35.93	15.58	32.99	14.74	42.16	<u>17.10</u>	74.48	66.19	34.49
CatSGG	91.63	7.52	45.81	<u>4.32</u>	43.09	49.23	37.74	7.26	92.10	89.78	<u>42.08</u>
CatSGG+	91.63	<u>8.60</u>	<u>44.90</u>	0.00	<u>42.86</u>	<u>46.04</u>	39.32	22.55	<u>92.08</u>	89.78	43.11

tool-anatomy relationships is beneficial for recognizing workflow stages that often hinge upon the correct sequence of instrument interactions. Notably, LLMs performed competitively in our experiments without temporality, but were not able to leverage temporal context or groundings for improved performance, despite being much larger models.

Table 4. Surgical Workflow Recognition Results

	Temporal Window	Semantic Relations	Spatial Encoding	Accuracy	F1
Holm et al. [10]	1 frame	✗ CATARACTS	✓	65.56	52.24
GATv2	1 frame	✓ CAT-SG	✓	70.81	56.02
Llama 3.2 3B	1 frame	✓ CAT-SG	✗	67.70	53.98
Llama 3.2 3B	1 frame	✓ CAT-SG	✓	69.13	53.87
Holm et al. [10]	30 frames (90 s)	✗ CATARACTS	✓	73.77	64.93
GATv2	30 frames (90 s)	✓ CAT-SG	✓	78.63	70.15
Llama 3.2 3B	30 frames (90 s)	✓ CAT-SG	✗	29.02	5.52

Surgical Technique Recognition. Table 5 shows that our GATv2 model, utilizing temporal graphs sampled at 5 fps, achieves the highest accuracy and F1 scores. Interestingly, longer temporal windows sampled with lower resolution (50 s at 1 fps) result in lower accuracy, suggesting that the fine-grained instrument motions crucial for distinguishing specific techniques occur within shorter time spans. This highlights the value of our dataset being annotated at 5 fps. Furthermore, ablating spatial features degrades performance, confirming that precise tool positioning is vital for differentiating lens-breaking strategies.

Table 5. Surgical Technique Recognition Results

	Temporal Window (sampled fps)	Spatial Features	Accuracy	F1
GATv2	10 s (5 fps)	✓	68.75 ± 4.11	48.40 ± 1.72
GATv2	50 s (1 fps)	✓	66.34 ± 1.38	41.31 ± 1.83
GATv2	10 s (5 fps)	✗	64.48 ± 1.15	44.52 ± 1.58

6 Conclusion

In this paper, we introduce CAT-SG, a large, fine-grained scene graph dataset specifically designed to understand complex workflows in cataract surgery. By capturing not only the presence of surgical instruments and anatomical structures but also their intricate interactions, CAT-SG provides an unprecedented level of detail for modeling real-world procedural variations. Alongside the dataset, we propose CatSGG, a novel scene graph generation model that leverages pre-trained video backbones and efficiently processes both spatial and temporal information. The holistic view offered by CAT-SG has the potential to enhance a wide range of applications, from automated skill assessment and decision support to advanced surgical training. Furthermore, this structured representation of surgical procedures promotes more transparent and explainable approaches, as each relationship can be precisely interpreted in a clinical context. Ultimately, this can improve patient outcomes and support surgeons in delivering safer, more efficient care.

Acknowledgments. The authors would like to thank Carl Zeiss AG for support of this research.

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

References

1. Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., Sahu, M., Galdran, A., Araújo, T., Vo, D.M., Panda, C., Dahiya, N., Kondo, S., Bian, Z., Vahdat, A., Bialopetravičius, J., Flouty, E., Qiu, C., Dill, S., Mukhopadhyay, A., Costa, P., Aresta, G., Ramamurthy, S., Lee, S.W., Campilho, A., Zachow, S., Xia, S., Conjeti, S., Stoyanov, D., Armaitis, J., Heng, P.A., Macready, W.G., Cochener, B., Quélélec, G.: Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis* **52**, 24–41 (2019). <https://doi.org/https://doi.org/10.1016/j.media.2018.11.008>, <https://www.sciencedirect.com/science/article/pii/S136184151830865X>
2. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=F72ximsx7C1>
3. Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: Valor: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345 (2023)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
5. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 (2023)

6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
7. Funke, I., Rivoir, D., Speidel, S.: Metrics matter in surgical phase recognition (2023), <https://arxiv.org/abs/2305.13961>
8. Gastager, D., Ghazaei, G., Patsch, C.: Watch and learn: Leveraging expert knowledge and language for surgical video understanding. ArXiv **abs/2503.11392** (2025), <https://api.semanticscholar.org/CorpusID:277044026>
9. Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D.: Cadis: Cataract dataset for image segmentation. arXiv preprint arXiv:1906.11586 (2019)
10. Holm, F., Ghazaei, G., Czempel, T., Özsoy, E., Saur, S., Navab, N.: Dynamic scene graph representation for surgical video. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 81–87 (2023)
11. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10236–10247 (2020)
12. Köksal, Ç., Ghazaei, G., Holm, F., Farshad, A., Navab, N.: Sangria: Surgical video scene graph optimization for surgical workflow prediction. arXiv preprint arXiv:2407.20214 (2024)
13. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
14. Luo, Z., Durante, Z., Li, L., Xie, W., Liu, R., Jin, E., Huang, Z., Li, L.Y., Wu, J., Niebles, J.C., Adeli, E., Fei-Fei, L.: MOMA-LRG: Language-refined graphs for multi-object multi-actor activity parsing. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=eJhc_CPXQIT
15. 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)
16. Rodin, I., Furnari, A., Min, K., Tripathi, S., Farinella, G.M.: Action scene graphs for long-form understanding of egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18622–18632 (2024)
17. Wang, G., Li, Z., Chen, Q., Liu, Y.: Oed: towards one-stage end-to-end dynamic scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27938–27947 (2024)
18. Wang, J., Wen, Z., Li, X., Guo, Z., Yang, J., Liu, Z.: Pair then relation: Pair-net for panoptic scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
19. Yang, J., Ang, Y.Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z.: Panoptic scene graph generation. In: European Conference on Computer Vision. pp. 178–196. Springer (2022)
20. Yang, J., Peng, W., Li, X., Guo, Z., Chen, L., Li, B., Ma, Z., Zhou, K., Zhang, W., Loy, C.C., et al.: Panoptic video scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18675–18685 (2023)

21. Zhou, Z., Shi, M., Caesar, H.: Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21637–21648 (2023)
22. Zhu, F., Yang, J., Jiang, H.: Towards flexible visual relationship segmentation. *Advances in Neural Information Processing Systems* **37**, 107633–107661 (2025)
23. Özsoy, E., Pellegrini, C., Keicher, M., Navab, N.: Oracle: Large vision-language models for knowledge-guided holistic or domain modeling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 455–465. Springer (2024)