Exploring Temporal Surgical-VQA Using Scene Graphs

Tiffany Marcus, M Shifat Hossain, Laura Brattain
University of Central Florida, Orlando, FL

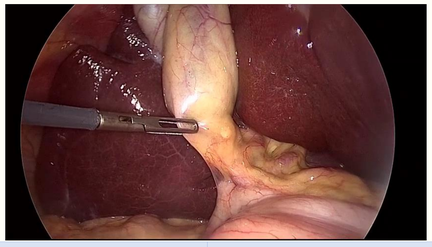
Center for Research in Computer Vision

Institute of Artificial Intelligence
UNIVERSITY OF CENTRAL FLORIDA

Introduction

Can temporal questions be answered using a scene-graph as a surgical scene representation?

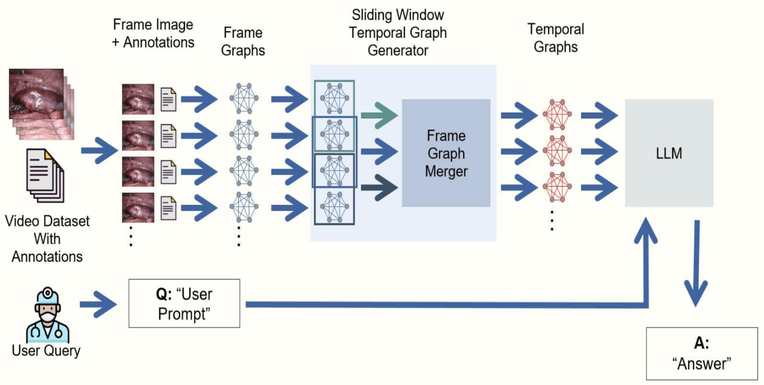
Scene-graphs have proven to be effective representation in surgical VQA[1].These studies have only evaluated static scenes. We aimed to answer questions pertaining to actions across multiple frames.



Which tool is operating on gallbladder	Grasper
What anatomy is at the top-left of the frame	Abdominal wall cavity
Which target is being grasped in the scene?	Gallbladder

Materials and Methods

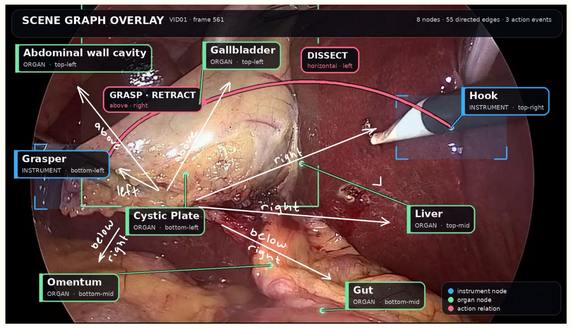
- 45 videos[2], one frame per second, average of 1,200 frames per video
- Contains annotated action triplets (<instrument, verb, target>) for every frame
- Bounding boxes denoting object location[1]



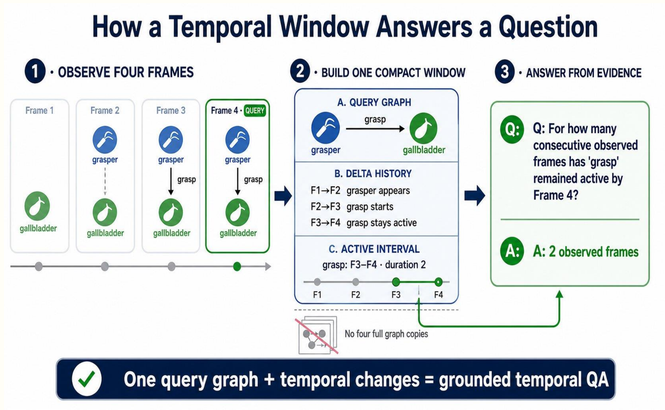
Temporal scene-graph question-answering pipeline. Annotated surgical video frames are converted into per-frame scene graphs and organized into overlapping temporal windows. The serialized temporal context and user query are then processed by a LoRA-adapted Qwen2.5-0.5B language model to generate an answer.

Experimental Results

Visualized scene-graph showing all action edges and only spatial edges from cystic plate



"Temporal windows" are created with four sequential frames where the changes between frames are recorded.



Non-temporal question accuracy	Temporal question accuracy
73.6%	99.7%

✓ Which instruments or organs are present?
✓ How many objects satisfy a spatial or action relationship?
✓ Does a particular object or relationship exist?
✓ What is the name or type of a referenced object?
✓ What action is an instrument performing?

✓ When did an action begin?
✓ How long has an action persisted?
✓ What action or state changed?
✓ Is the duration at least N observations?

Conclusion

- This study investigated whether temporal scene-graph representations can support surgical visual question answering
- Annotated video frames were converted into individual scene graphs and organized into overlapping four-observation windows that preserved the sequence of recorded surgical actions.
- The model achieved 99.7% exact-match accuracy on template-generated temporal questions.
- Demonstrated the feasibility of using compact temporal scene-graph context for structured surgical QA.
- The temporal questions were template-generated, limiting the diversity and difficulty of the evaluation
- Questions were created to examine short-term changes

Future Work

- More diverse questions, longer temporal windows, and temporal ablations to assess generalization to more challenging reasoning tasks.
- Compare Qwen2.5-0.5B with larger Qwen variants using the same training and evaluation splits.
- Expand non-temporal training to the complete available QA dataset and measure performance as data volume increases.
- Explore adaptive windows for longer surgical actions without substantially increasing graph size.

Acknowledgements

- UCF Computer Vision REU funded by the NSF
- Contributions by members of the UCF Department of Computer Science and the UCF College of Medicine

References

[1] Yuan, K., Kattel, M., Lavanchy, J. L., Navab, N., Srivastav, V., & Padoy, N. (2023). Advancing Surgical VQA with Scene Graph Knowledge. In arXiv.org. https://arxiv.org/abs/2312.10251

[2] Nwoye, C. I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., & Padoy, N. (2022). Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis, 78(1361–8415), 102433. https://doi.org/10.1016/j.media.2022.102433

[3] Drago, M. O., Luca, C., Celebi, B. P., Pierantozzi, D., Chiara, L., Cesare, H., Stoyanov, D., De Momi, E., Bano, S., & Mobarak I, H. (2025). SurgViVQA: Temporally-Grounded video question answering for Surgical scene understanding. arXiv.Org. https://arxiv.org/abs/2511.03325