Florence-Lapchole: Automated Surgical Scene Inference and Segmentation with Natural Language-Guided Online Interface

Jonathan A Hubermann, Marzie Lafouti, Michelle Levy, Dr. Liane S Feldman, Dr. Amir Hooshiar

Surgical Performance Enhancement and Robotics Centre (SuPER), Department of Surgery, McGill University

Introduction

Laparoscopic surgical videos, such as laparoscopic cholecystectomy, offer valuable insights for skill assessment and tool-tissue interaction. Advances in deep learning models have improved object detection and tracking but remain untested on surgical videos. This study presents an online platform for automatic segmentation and annotation, reducing manual effort and enhancing AI training.

Methods and Procedures

User Input and Natural Language Processing

1. User uploads a surgical video and types prompt specifying the object or tissue of interest.
2. GPT-4o-mini model (OpenAI, CA, USA) processes and returns a single-word classification for one of 11 predefined classes, i.e., abdominal wall, liver, fat, grasper, blood, liver ligament, connective tissue, cystic duct, gallbladder, gastrointestinal tract, and hook.

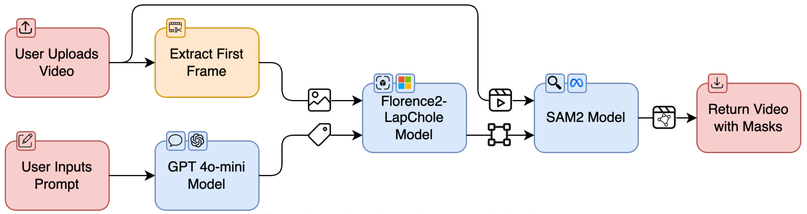


Fig. 1: Complete flow of automated pipeline

Preliminary Object Detection with Florence2-LapChole

The classification and first video frame are input to the Florence2-LapChole[1] model (a fine-tuned version of Microsoft's Florence2[2]). This image-text-to-text model performs object detection to identify the requested tools or structures, and returns delineating bounding boxes.

Segmentation Using SAM2

The bounding box coordinates and full surgical video are fed into SAM2 (Segment Anything Model 2 by Meta[3], CA, USA) for segmentation. SAM2 first segments selected frames for user verification, allowing adjustments if needed. Once confirmed, segmentation extends to the entire video, ensuring object masks remain consistent even when objects exit or re-enter the frame.

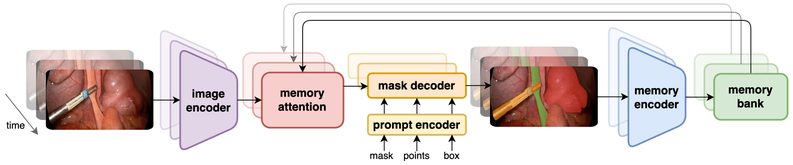


Fig. 2: The SAM2 architecture

Output and Interactive Web Interface

The pipeline outputs an annotated video with masks overlaid on each frame. An interactive cloud-based web interface (Fig. 3) provides easy access for clinicians and researchers.

Results

The online interface, built in TypeScript, is integrated with a Python-based serverless backend that forwards requests to a persistent GPU-enabled virtual machine for real-time processing, hosted by Amazon Web Services (WA, USA).

- Florence2-LapChole achieved an accuracy of 82% using the Intersection Over Union (IoU) metric.
- Average time for segmentation of 60-second videos (1827 frames) was 113.2-s, exhibiting a computation time of 61.96-ms per frame.
- Total of 6.2 minutes from the user's input to delivery of the final segmented video.

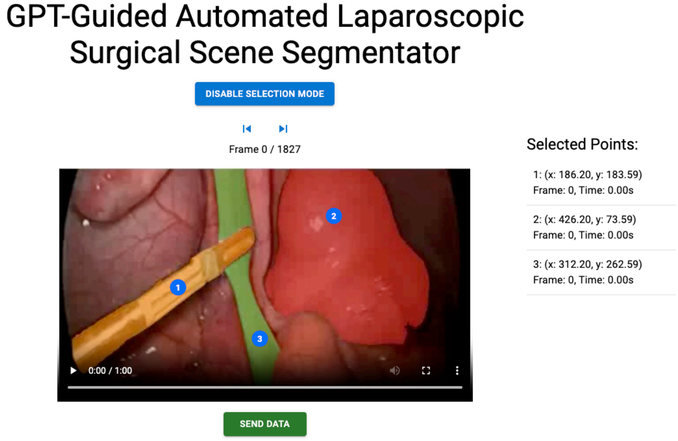


Fig. 3: Three auto-generated masks on a representative LapChole video

Conclusion

This AI-powered platform accelerates laparoscopic video annotation and segmentation from hours to minutes, enabling rapid generation of robust datasets. Its ease-of-access and ability to handle complex object movements and occlusions make it a crucial resource for surgical training, skill assessment, and future AI developments.

References

1. Levy M, Lafouti M, Feldman L and Hooshiar A (2024). Florence-Lap: A Real-time Tissue Segmentation and Phrase Grounding for Laparoscopic Surgical Video Analysis, Alberta Robotics & Intelligent Systems Expo (ARISE 2024).
2. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., … Yuan, L. (2023). Florence-2: Advancing a unified representation for a variety of vision tasks. doi:10.48550/ARXIV.2311.06242
3. Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., … Feichtenhofer, C. (2024). SAM 2: Segment Anything in Images and Videos. doi:10.48550/ARXIV.2408.00714