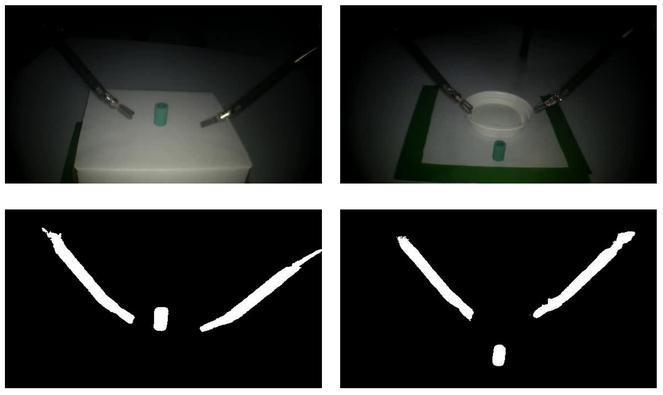
Real-Time Video Segmentation for Autonomous Robotic Manipulation

Chetan Reddy Narayanaswamy
chetanrn@stanford.edu

Vakula Venkatesh
vakulav@stanford.edu

Introduction

- **Motivation:**
  Enable real-time scene understanding for autonomous robotic surgical manipulation tasks.

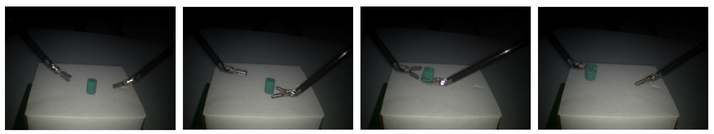


But why do we need Segmentation Masks?
They improve generalization in manipulation tasks by remaining consistent across backgrounds.

- **Challenge:**
  High-accuracy models (e.g., SAM2) are too slow for 30 Hz closed-loop control
- **Objective:**
  Mimic SAM2-quality masks using a fast, lightweight U-Net for real-time inference

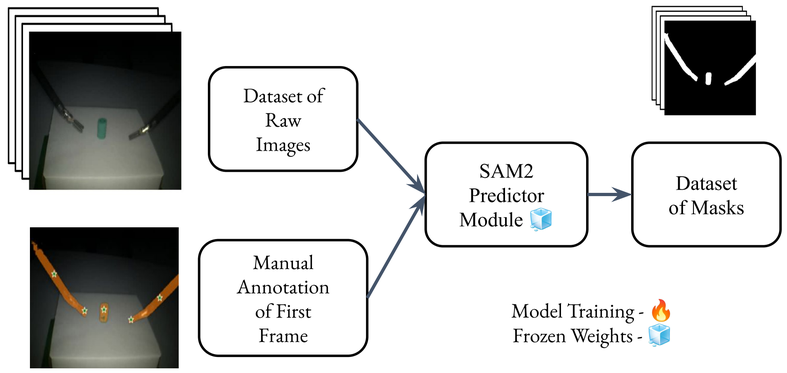
Dataset

- **Platform:** daVinci Surgical Robot
- **Data Collected:** RGB frames + joint angles
- **Frame Rate:** Recorded at 30Hz
- **Demonstration:** Object transfer task

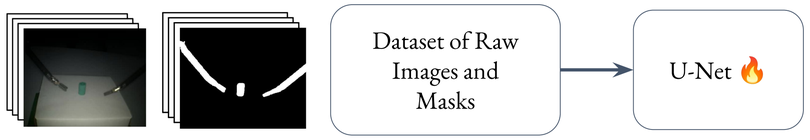


Methodology

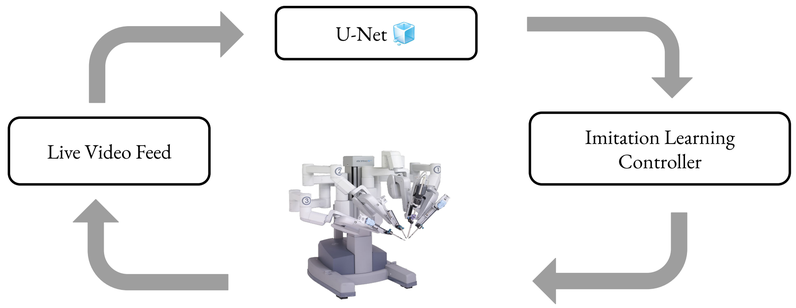
- **SAM2 for Pseudo-Ground-Truth Generation:**
  - Manual annotation on a single frame to provide the segmentation Region of Interest
  - Propagate it through the video to generate masked training data



- **Lightweight U-Net for Real-Time Segmentation:**
  - Encoder-decoder with skip connections
  - 2 downsampling blocks → Bottleneck→ 2 upsampling blocks
  - Binary Cross-Entropy Loss (BCE)

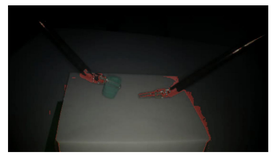


- **Trained U-Net Segmentation Module in the Controller Loop**



Baselines

- Classical Computer Vision based Masking

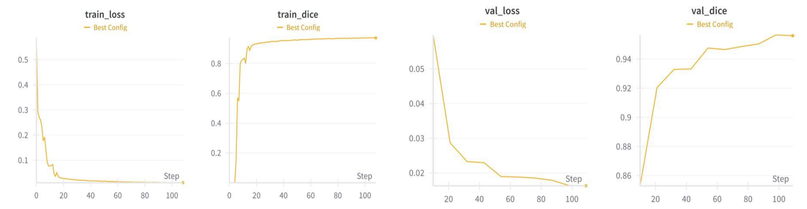


- Textual Prompt-Based Masking using CLIPSeg

"surgical robot" "robotic arm"

Results

- **Training & Validation Curves**



Model	Dice (%)	Inference Time (ms)	Maximum FPS
SAM2	100	600	1.6
U-Net (GPU)	**95.6**	**10**	**100**
U-Net (CPU)	95.6	80	12.5
CLIPSeg	81.3	150	6.6
Classical CV	63.2	5	200

- **U-Net** achieves **95.6% Dice score** at a real-time speed of 30 Hz
- **60x speedup** over SAM2 with minimal drop in accuracy

References

1. Ravi, Nikhila, et al. "Sam 2: Segment anything in images and videos." *arXiv preprint arXiv:2408.00714* (2024).
2. Qiu, Liang et al. "Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network." Healthcare technology letters vol. 6,6 159-164. 5 Dec. 2019, doi:10.1049/htl.2019.0068
3. Siddique, Nahian, et al. "U-Net and its variants for medical image segmentation: theory and applications." arXiv preprint arXiv:2011.01118 (2020).