SURGVU 2026 · WEEK 6 PRESENTATION

Surgical Visual Understanding 2026 Category 1 Challenge

DeAndre Bailey

Project Recap

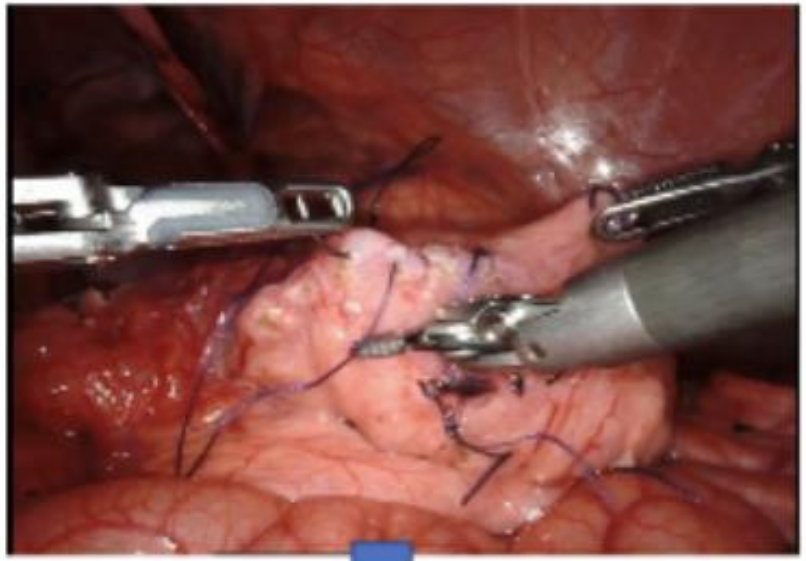
- Surgical Visual Understanding 2026 Category 1 Challenge
- Classify and localize surgical instruments using bounding boxes
- Develop **semi-supervised or weakly supervised models**



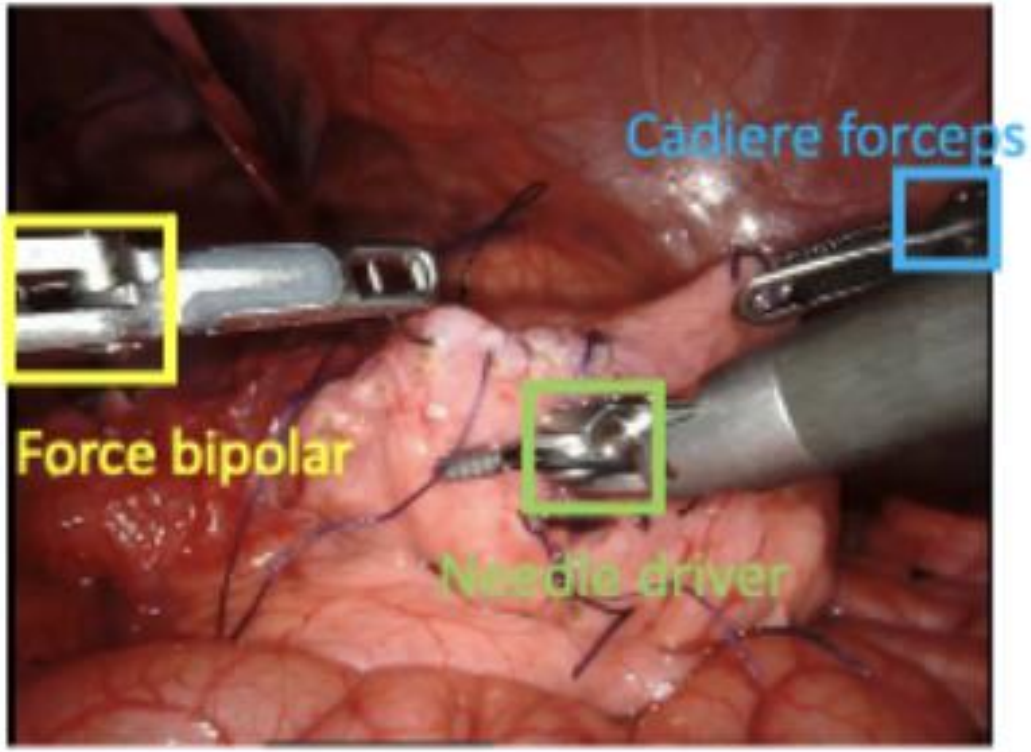
Project Recap

Category 1
(weak/noisy labels)

Input image



Classify and localize tools



Extreme Class Imbalance

- 14 Tool Classes Total, 12 Tool Classes are tested on
- Limited to no ground truths for some classes

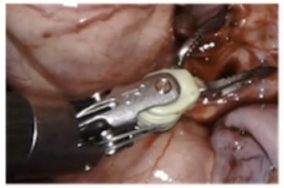

Needle driver


Monopolar curved scissors

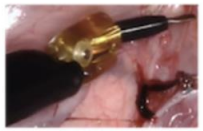

Force bipolar


Clip applier


Cadiere forceps


Bipolar forceps


Vessel sealer


Permanent cautery hook/spatula


Prograsp forceps

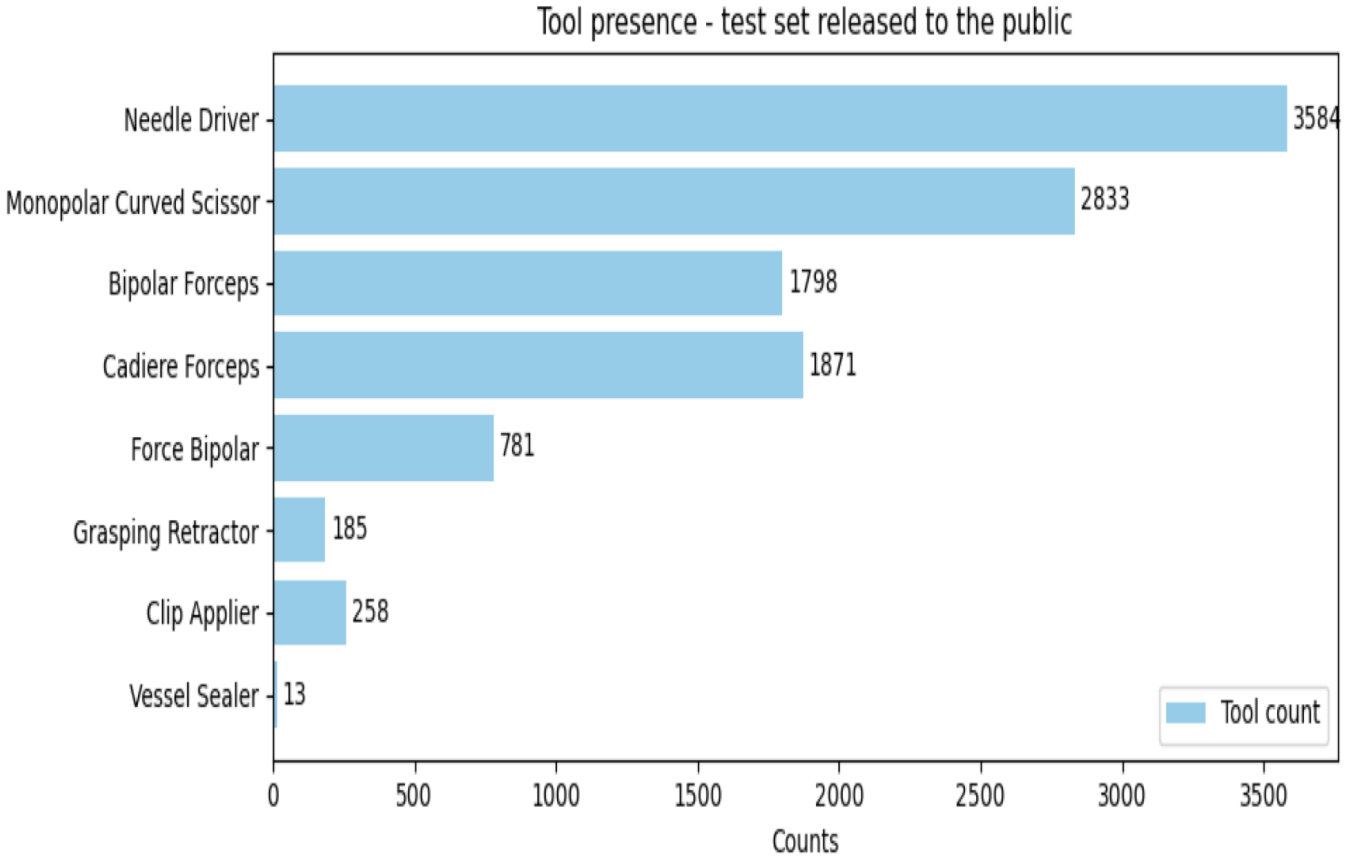

Stapler


Grasping retractor


Tip-up fenestrated grasper

Extreme Class Imbalance

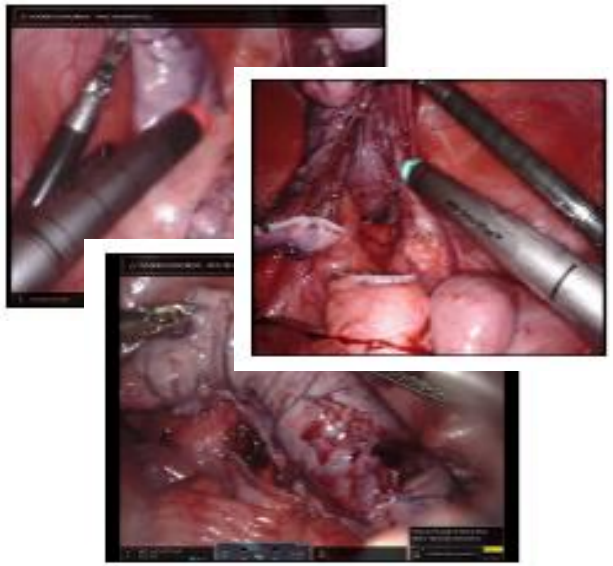
- Ground Truths for Only 8 Tools
- How do we get more annotations?



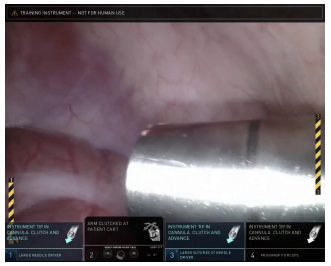
Data Annotation

Frame Selection

Use **K-Center Greedy**

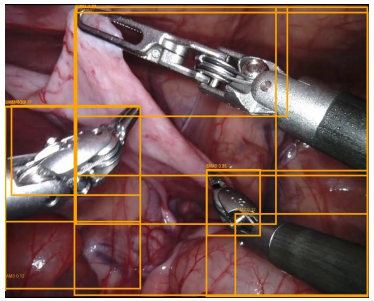


Filter and Annotate





Creating bounding boxes



Train Model

- Add Annotations to dataset
- Train the **Model** on the new dataset

5 classes went from 0 → 0.5–0.8 AP

Class	Baseline AP	New AP	Change
permanent cautery hook spatula	0.000	0.826	**+0.826**
vessel sealer	0.000	0.734	**+0.734**
suction irrigator	0.000	0.726	**+0.726**
prograsp forceps	0.000	0.656	**+0.656**
tip up fenestrated grasper	0.000	0.546	**+0.546**
stapler	0.000	0.035	**+0.035**

K-Center-Greedy

ACTIVE LEARNING FOR CONVOLUTIONAL NEURAL NETWORKS: A CORE-SET APPROACH

Ozan Sener*
Intel Labs
ozan.sener@intel.com

Silvio Savarese
Stanford University
ssilvio@stanford.edu

- How can we minimize annotation effort?
- Core-set selection approach
- Maximize visual diversity
  - Choosing Frames furthest from each other
- We use **RF-DETR DINOv2 backbone** to extract feature distance

$$\min_{\mathbf{s}^1:|\mathbf{s}^1|\leq b} \max_{i} \min_{j\in\mathbf{s}^1\cup\mathbf{s}^0} \Delta(\mathbf{x}_i, \mathbf{x}_j)$$

K-Center-Greedy

ACTIVE LEARNING FOR CONVOLUTIONAL NEURAL NETWORKS: A CORE-SET APPROACH

Ozan Sener*
Intel Labs
ozan.sener@intel.com

Silvio Savarese
Stanford University
ssilvio@stanford.edu

$$\min_{\mathbf{s}^1:|\mathbf{s}^1|\leq b} \max_{i} \min_{j\in\mathbf{s}^1\cup\mathbf{s}^0} \Delta(\mathbf{x}_i,\mathbf{x}_j)$$

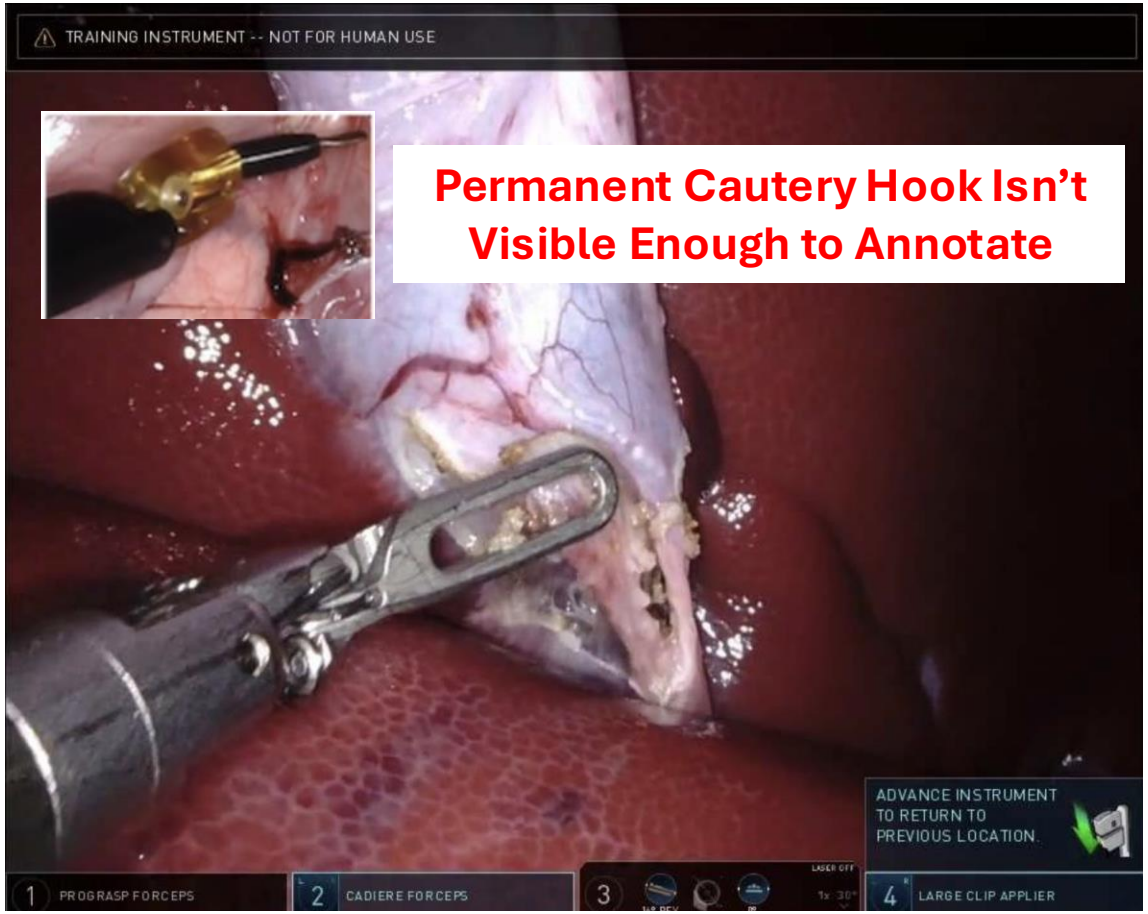
View *worst-case* frame Analyze Nearest Neighbors Calculates Distance

Optimize over all possible selections of the size b (annotation budget)

Manual Review: Discarded Frames



Prograsp Forceps – case_053 (Obscured)

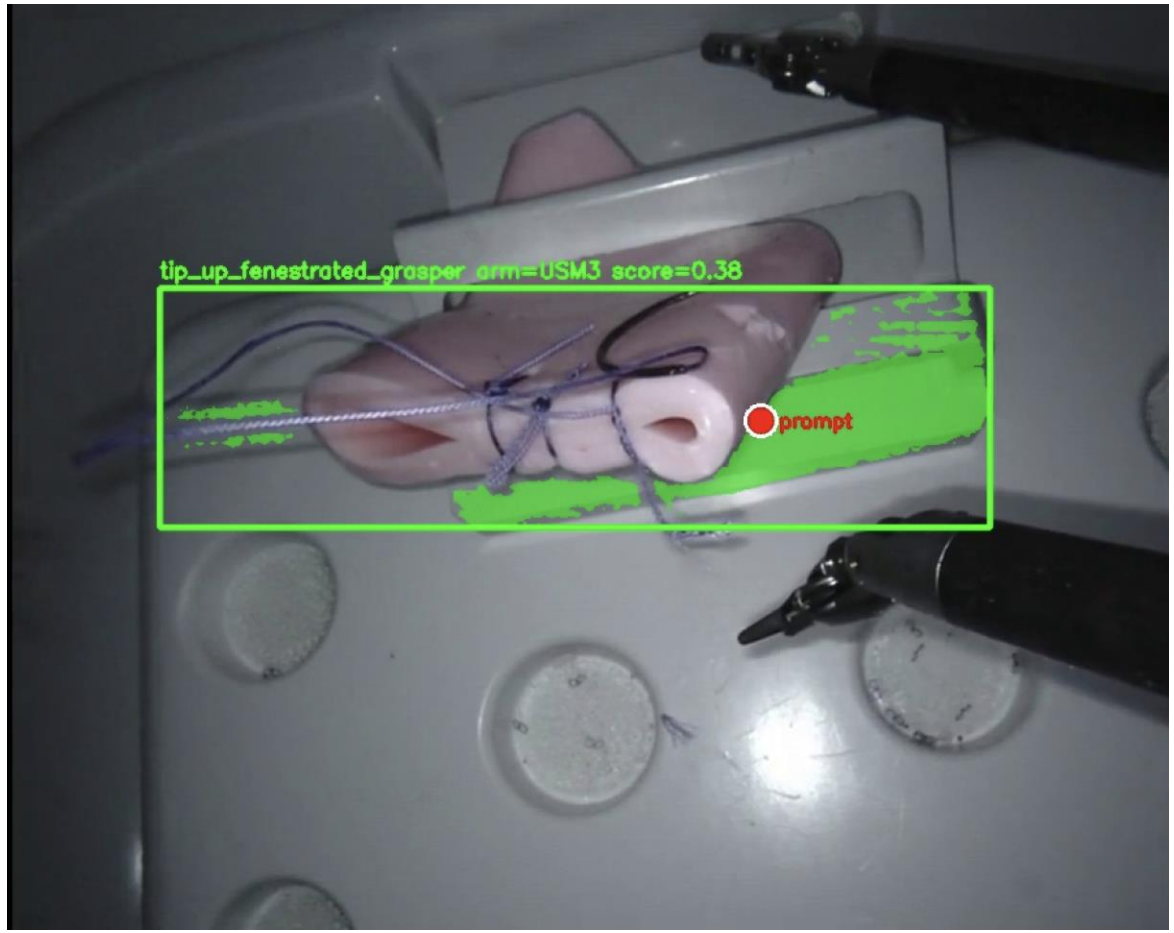


Permanent Cautery Hook — case_003 (Different Tool)

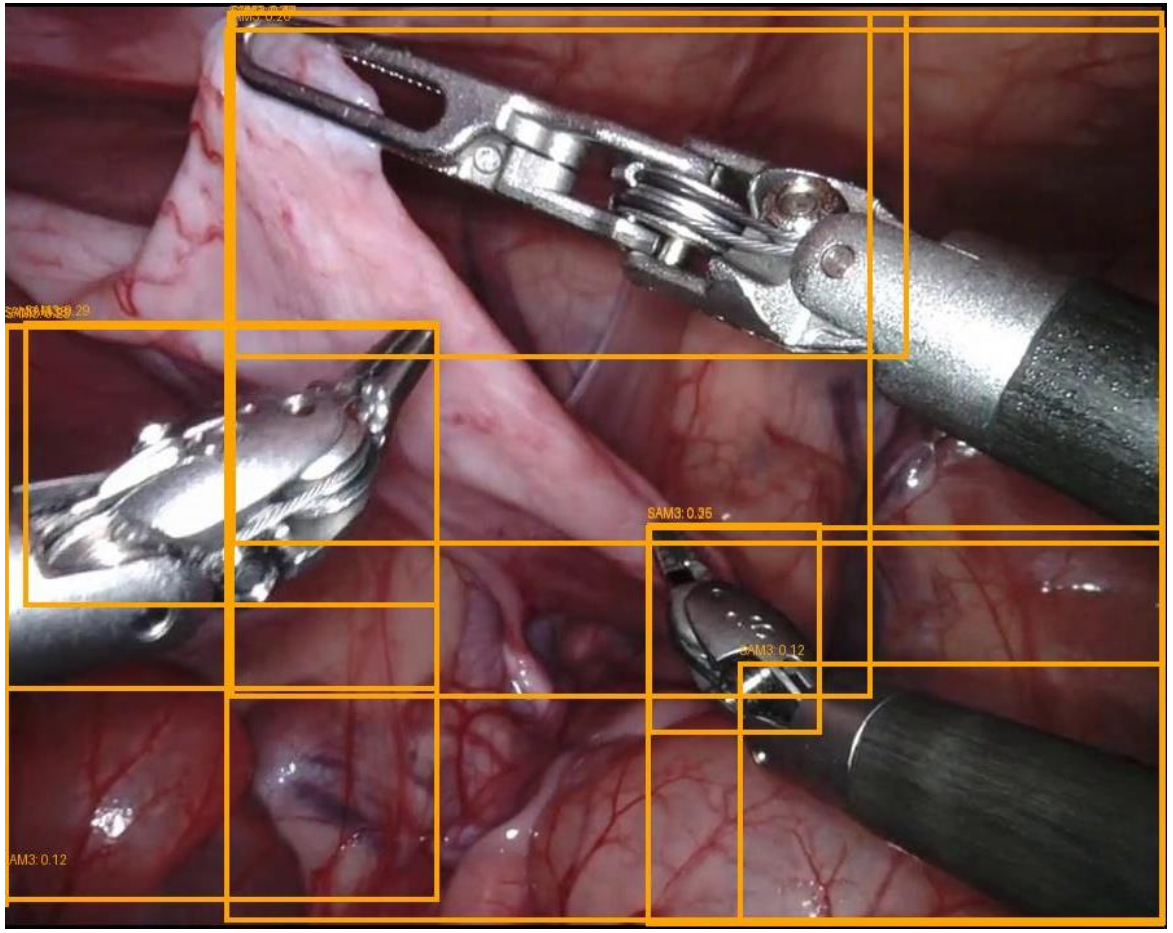
Bounding Boxes: SAM2, SAM3, and OwlV2

- **Text Prompting:**
  - 165 prompts Too many bounding boxes
- **Point Prompting:**
  - Worked well on rigid staplers
  - But, Failed on other tools
- Is it worth it to automize annotation process?
- Transitioned to high-quality manual annotation in Roboflow

Bounding Boxes: SAM2, SAM3, and OwlV2



Point Prompting (SAM2)

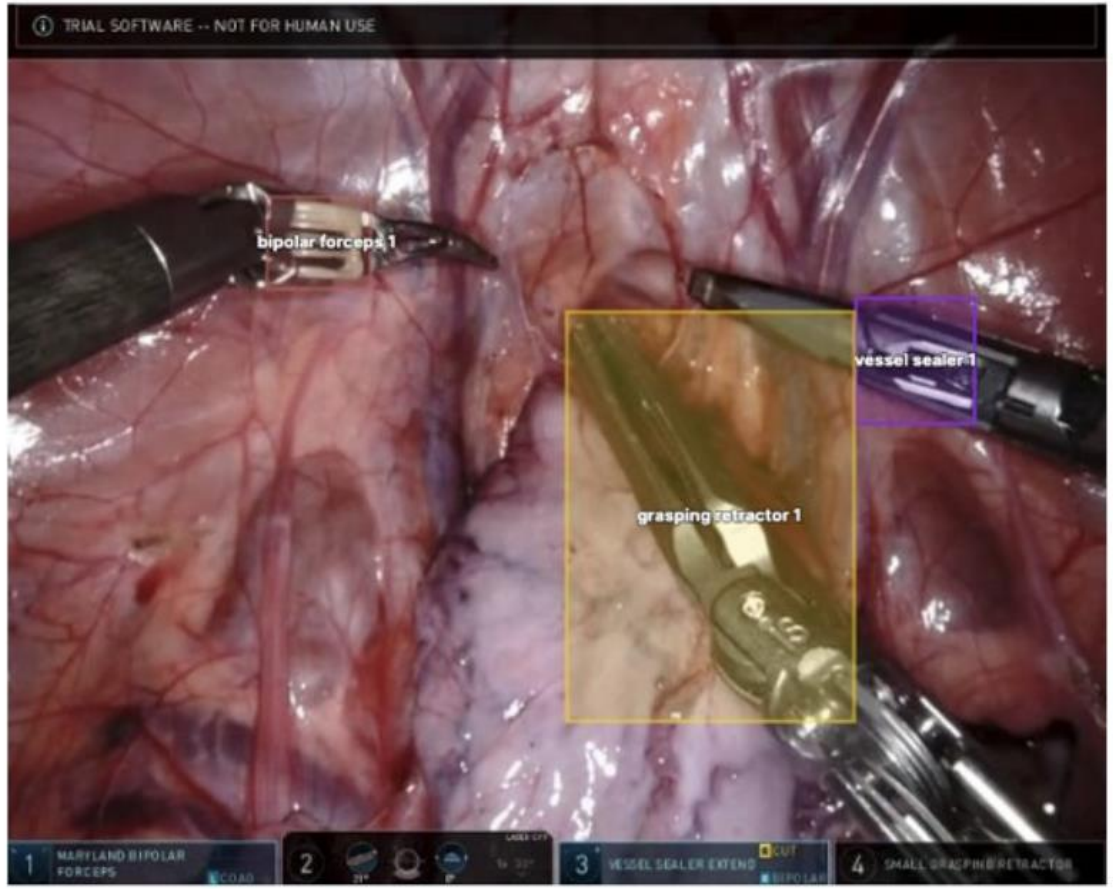


Text Prompting (SAM 3 and Owlv2)

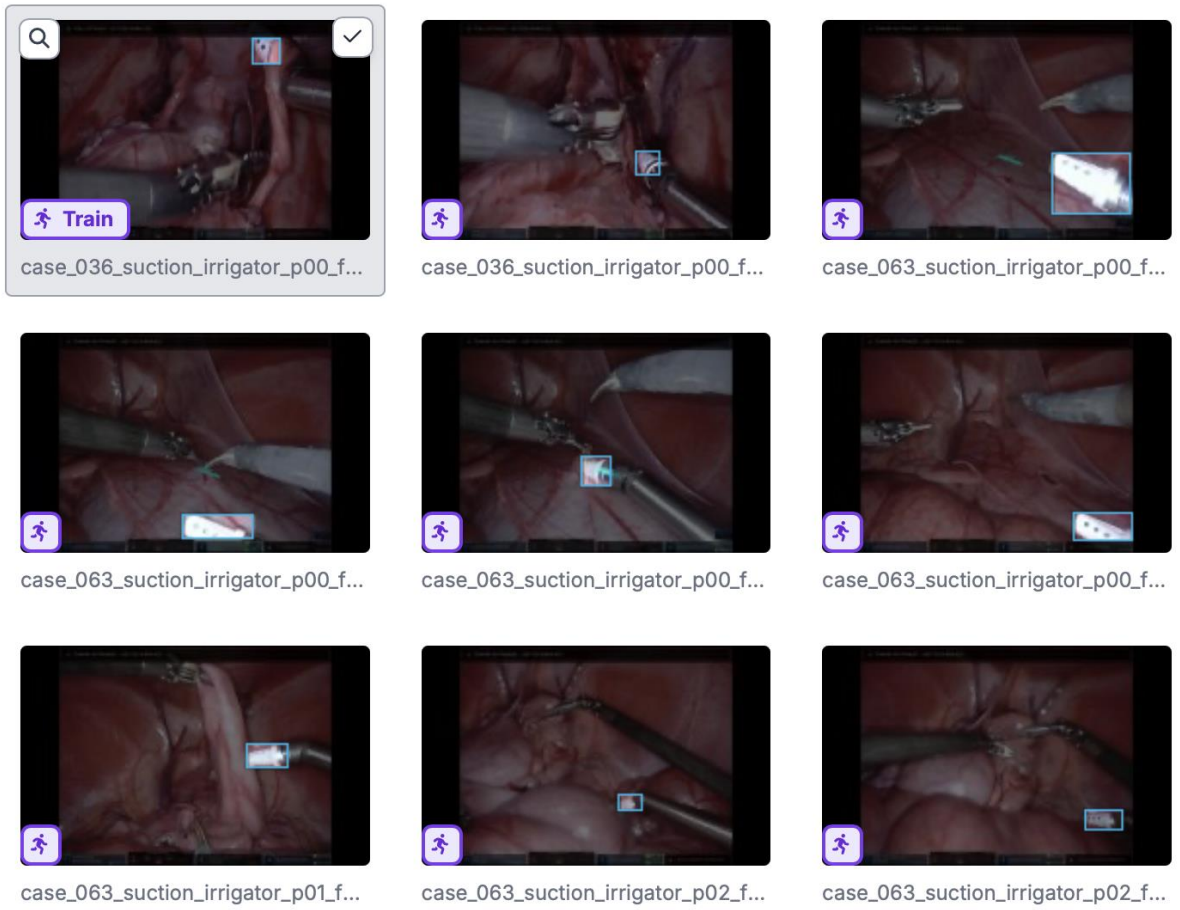
Bounding Boxes: Manual Annotation

- Manually Annotated **532 Endoscopic Frames**
  - Used Roboflow Annotation Tool
- Split and add to the training dataset
- Next: Train Model

Bounding Boxes: Manual Annotation

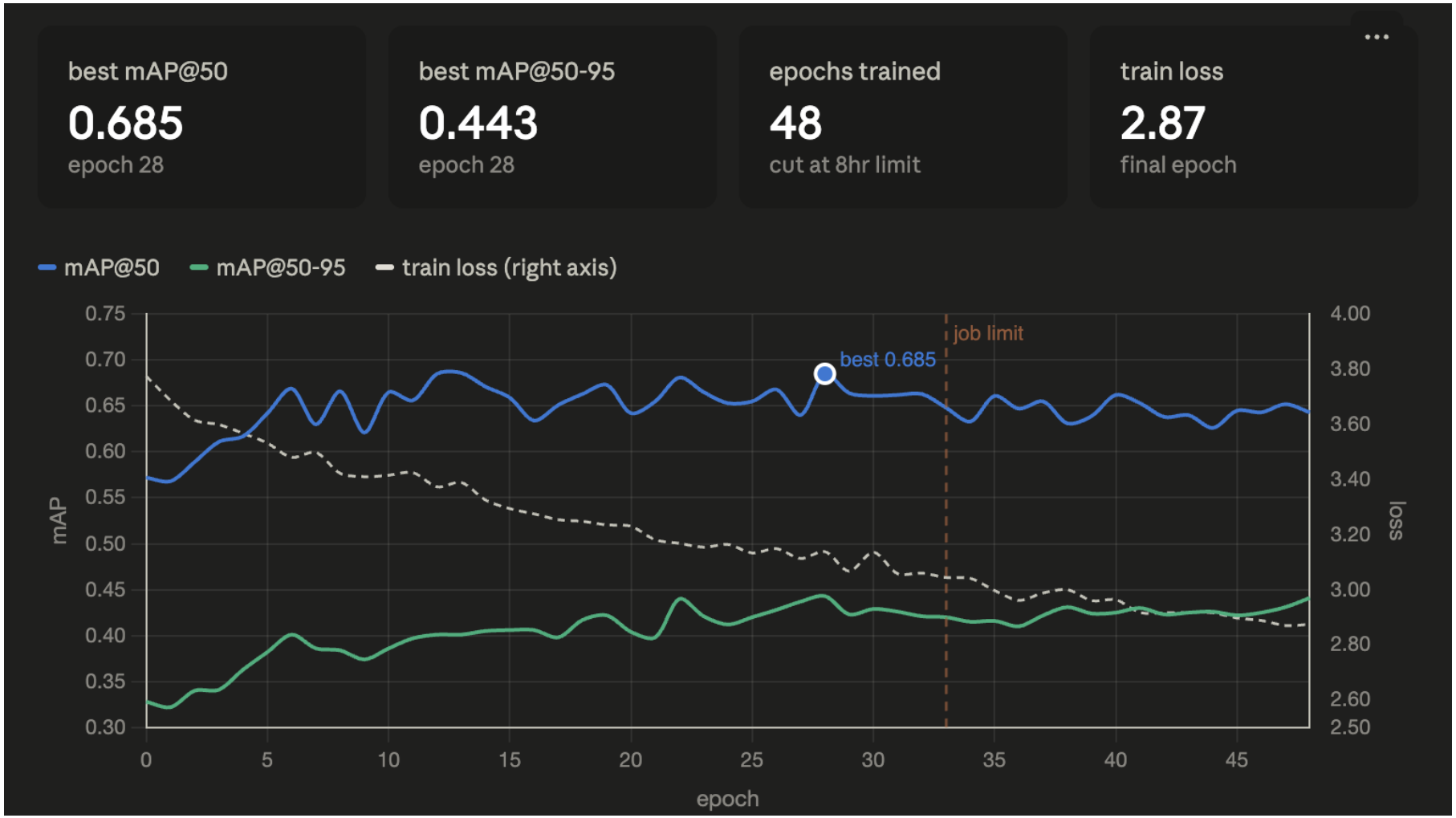


Annotation Examples Provided by Challenge



532 Frames Annotated Using Roboflow

RF-DETR Trained



Model Comparison

Evaluated on sessions 001 and 006. ViT-based architectures are superior

Model	mAP@50	mAP@50-95	Params
RF-DETR-L	***0.656***	***0.352***	*33.9M*
RT-DETR-L	0.576	0.271	32.0M
YOLOv11x			56.8M
YOLOv12l			26.3M
YOLOv8l	0.559	0.272	43.6M
YOLO26x	0.527	0.278	55.6M

RF-DETR Performs the best

Next Steps

- Identify weak classes that need more data
- Pseudo Labeling (RF-DETR 14 Class Model)
- Label Propagation (**SAM 3.1 Temporal Label Propagation**)
- Data Augmentation (Lighting, Masking)
- Find other Datasets

Questions?