NSF REU 2026
Week 3

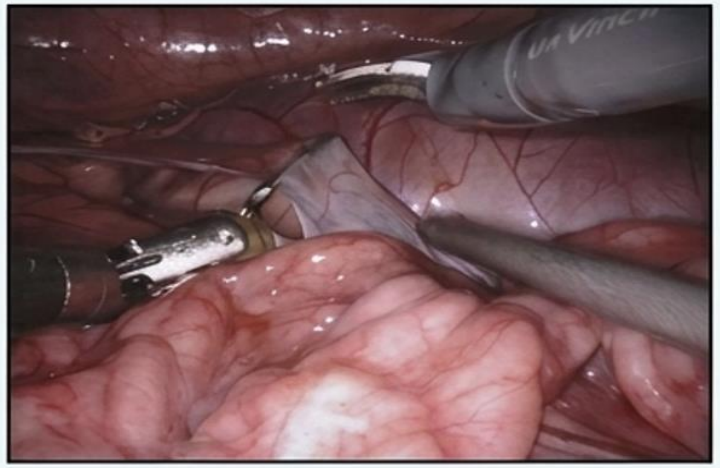
Tiffany Marcus

11 June 2026

The Clinical Need for Surgical VQA

- Expert surgeons are scarce and overloaded
- Limited time to educate trainees
- *Need to analyze surgical procedure performance and optimize the workflow*
- *Need to scale up the surgical training to produce high quality surgeons quickly*

Surgical VQA translates complex surgical scenes into easy-to-understand answers.



What tools are operating the organ?

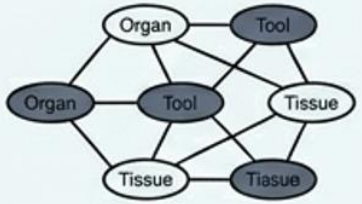
The tools operating are bipolar forceps, monopolar curved scissors, and suction.

The Gap in Surgical AI

Dimension 2: Output

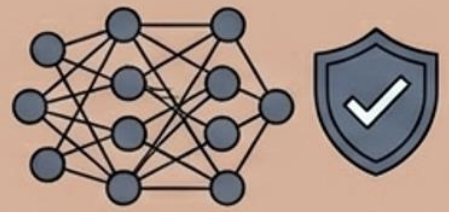
Dimension 1: Understanding

Scene Graphs + Black-Box



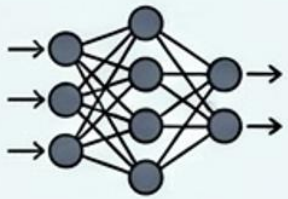
Models map objects relationally, but fail to communicate when unsure in overlapping surgical scenes.

Scene-Graphs + Calibrated Uncertainty



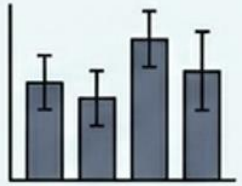
The Ideal Clinical VQA / This Proposal.
AI that parses scenes relationally and knows exactly when to say “I don’t know”.

Pixels + Black-Box



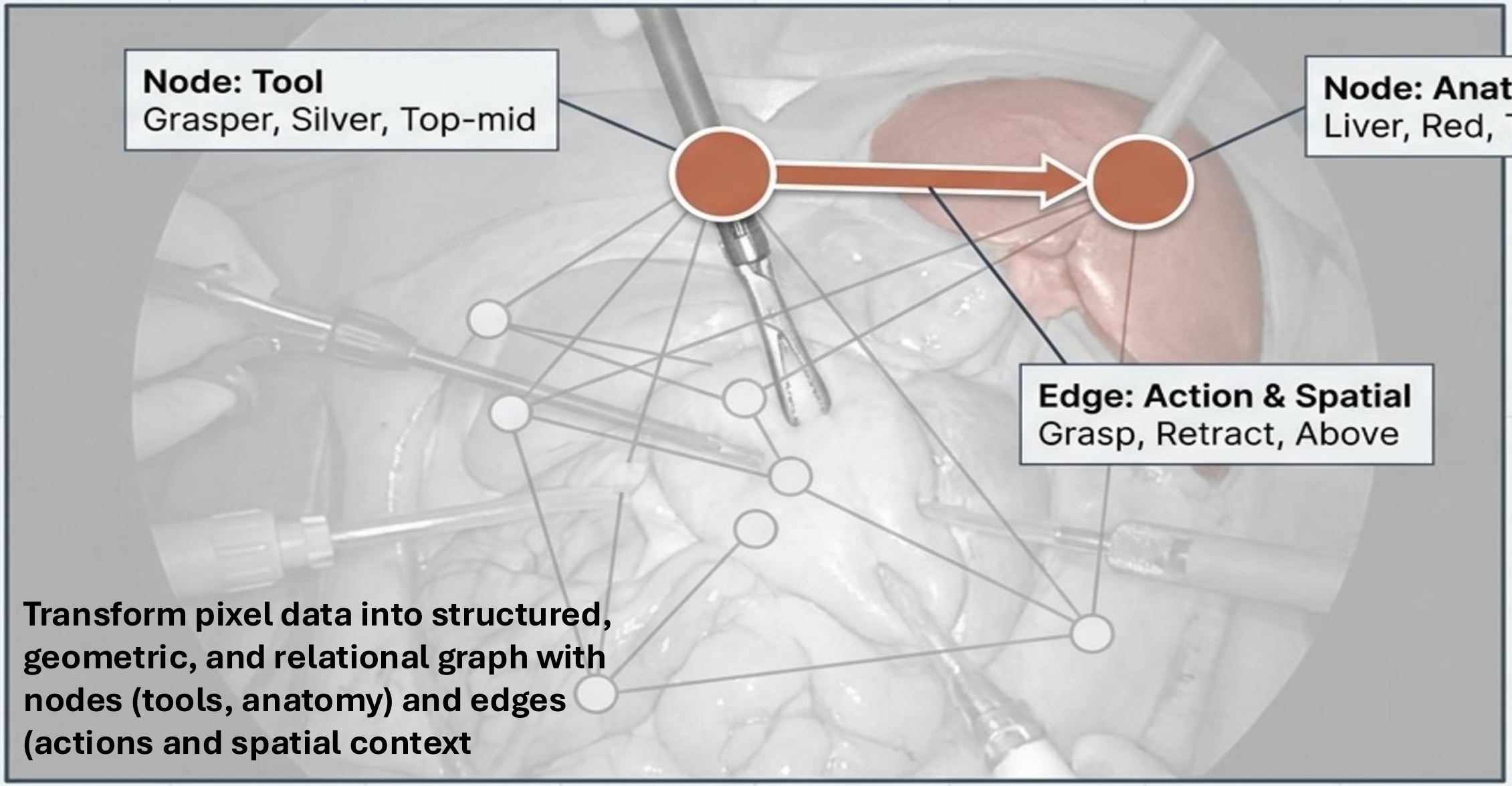
Standard models output raw labels from pixels but lack **relational logic** and hallucinate without warning.

Pixels + Calibrated Uncertainty



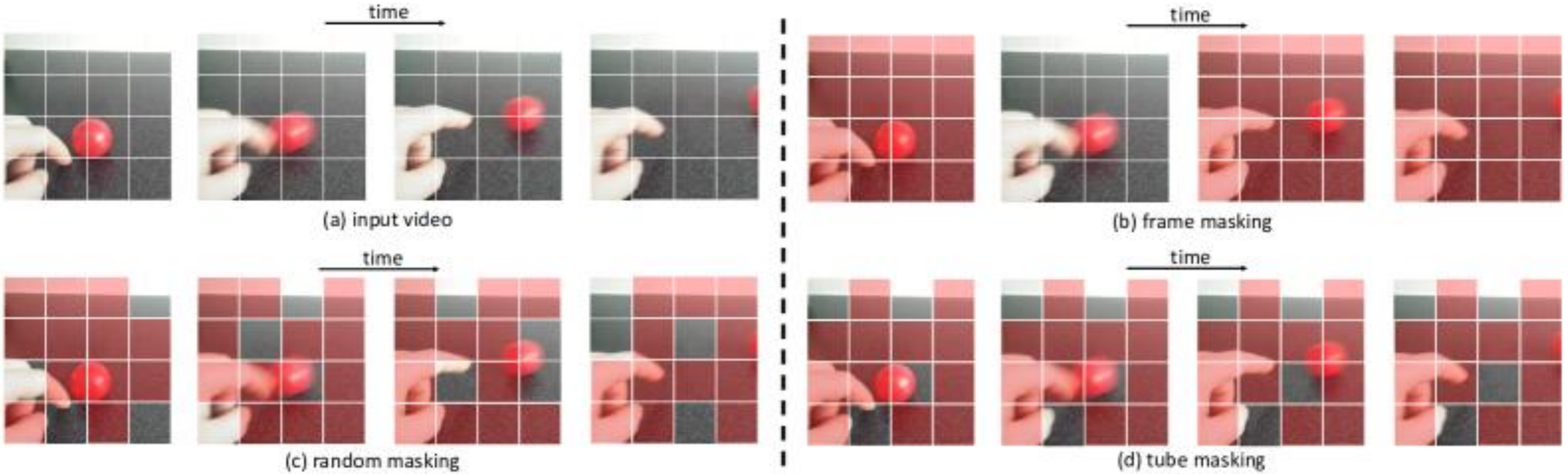
Models know when they are uncertain, but cannot answer complex, multi-hop geometric queries.

What is a Scene Graph



Transform pixel data into structured, geometric, and relational graph with nodes (tools, anatomy) and edges (actions and spatial context

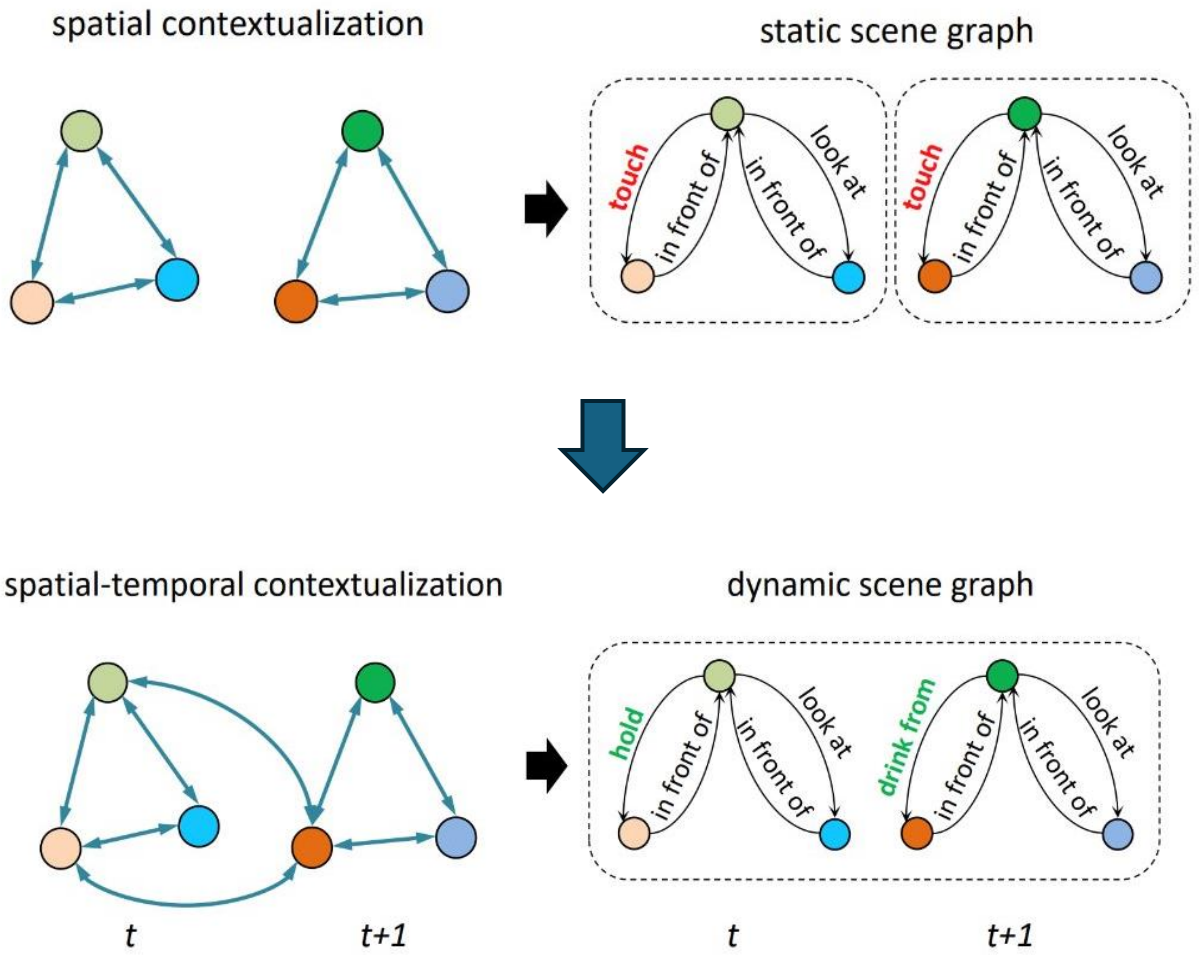
Temporal Model



Tong, Z., Song, Y., Wang, J., & Wang, L. (2022, October 18). *VideoMAE: Masked Autoencoders are data-efficient learners for self-supervised video pre-training*. arXiv.org. https://arxiv.org/abs/2203.12602

Dynamic Scene Graph

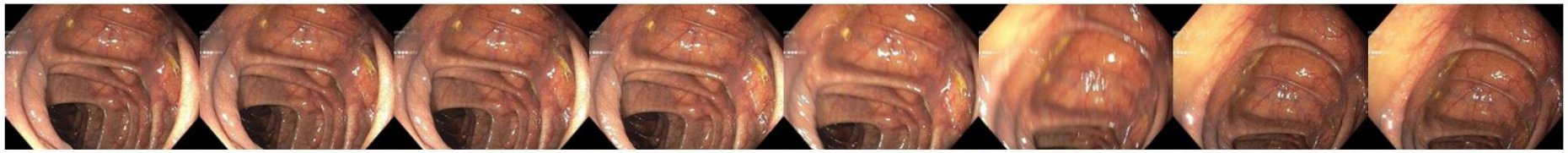




Spatial-Temporal Transformer for Dynamic Scene Graph Generation - GDLMA: Graph Deep Learning for Medical Applications - BayernCollab." *Dvb.bayern*, 2020, collab.dvb.bayern/spaces/TUMgdlma/pages/70348130/16.+Spatial-Temporal+Transformer+for+Dynamic+Scene+Graph+Generation. Accessed 11 June 2026.

Temporal and Spatial Questions

Temporal



QUESTIONS

In-Template: Is the scope primarily advancing or withdrawing in this sequence?

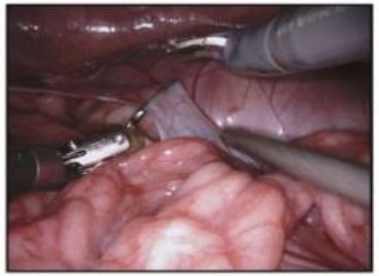
Out-Of-Template: Is the endoscope being advanced or retracted during this video?

ANSWERS

Long: Yes, the scope is withdrawing.

Short: yes

Static scene graph



Questions

Q1: what organ is being operated?
Q2: what is the state of bipolar forceps?
Q3: what is the state of suction?
Q4: where is bipolar forceps located?
Q5: what tools are operating the organ?

Answers

A1: kidney
A2: retraction
A3: tissue manipulation
A4: bipolar forceps is located at left-top
A5: the tools operating are bipolar forceps, monopolar curved scissors, and suction

Dynamic scene graph

Do we need to suction before retracting the bipolar forceps?

Next Steps

- Duplicate temporal surgical VQA (SurgViVQA)
- Add a scene graph layer (PyTorch Geometric Temporal Libray)
- Expand to larger datasets