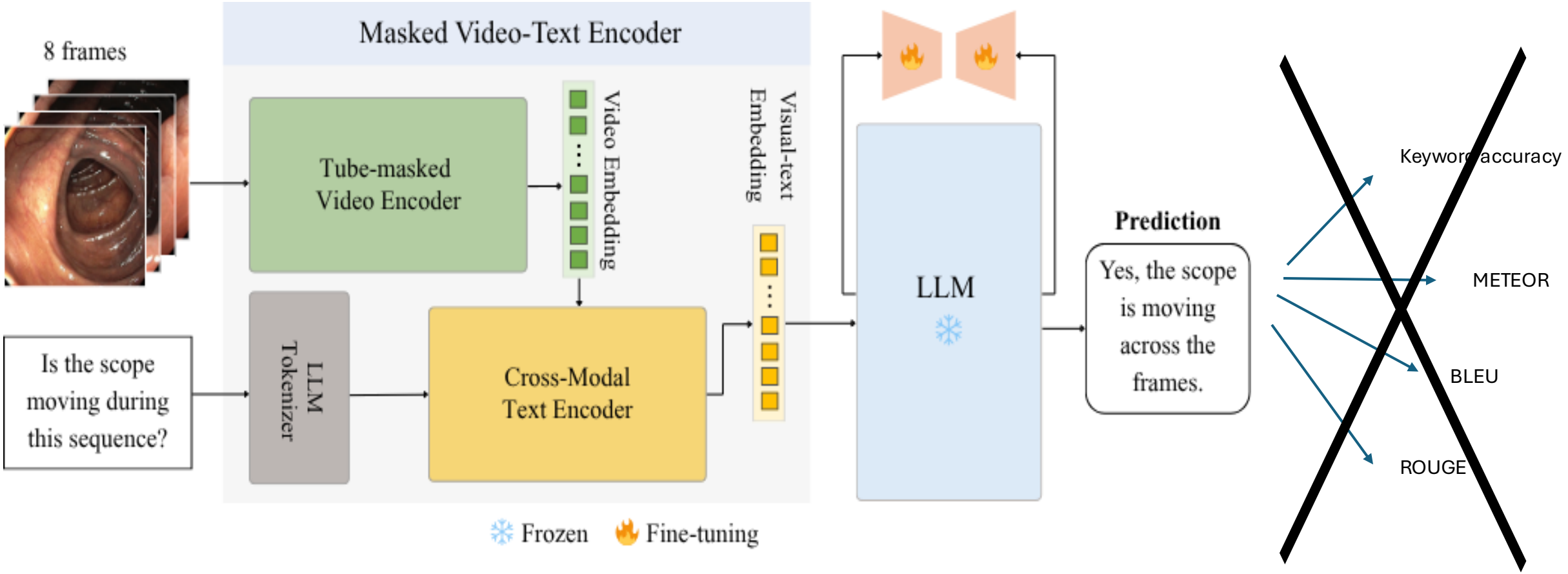
NSF REU 2026
Week 5

Tiffany Marcus

Mentor: Dr. Laura Brattain

25 June 2026

Cross Encoder-based Similarity Score



Mentor: Dr. Laura Brattain

Cross Encoder-based Similarity Score

- A cross encoder model is used to compare two texts
- The model used is - "nli-deberta-v3-small"
- NLI (natural Language Inference) models are used to evaluate a pair of texts, determining
  - Determining logical relationship
- The model determines if the pair is in –
  - Contradiction (Logically Impossible),
  - Entailment (Logical Truth)
  - Neutral (Logical Uncertainty)
- We considered two texts same if the pair is in "Entailment".
  - For other cases two texts are considered different.

CLIP Text Encoder Similarity Metric

- CLIP Text encoder is used
- Embedding vectors are computed for both reference and predicted texts
- Cosine similarity is computed between two embedding vectors
- The similarity score is considered as the metric for the two texts to be similar
- A cut-off threshold of 0.85 is set to decide if two inputs are similar

LLM-Based Similarity Scoring

- A small LLM is used as a judge for two inputs
- The model used is "Qwen2.5-0.5B-Instruct"
- The system prompt provided for the model is
  - "You are a strict, objective AI judge comparing two texts. Evaluate if the generated text matches the core meaning of the reference text. Output your evaluation purely as a JSON object with two keys:\n"
  - "1. 'qualitative': strictly 'Same' or 'Different'.\n"
  - "2. 'quantitative': integer from 0 to 100 representing the semantic match percentage.\n"
  - "Do not include any conversational text, markdown formatting, or explanations."
- The reference and generated texts are formatted as user inpus as follows -
  - "Reference Text: {reference_text}\n\nGenerated Text: {generated_text}"
- The LLM output is then parsed to get the score and normalized

Current Results on Test Set (Real-Colon-VQA)

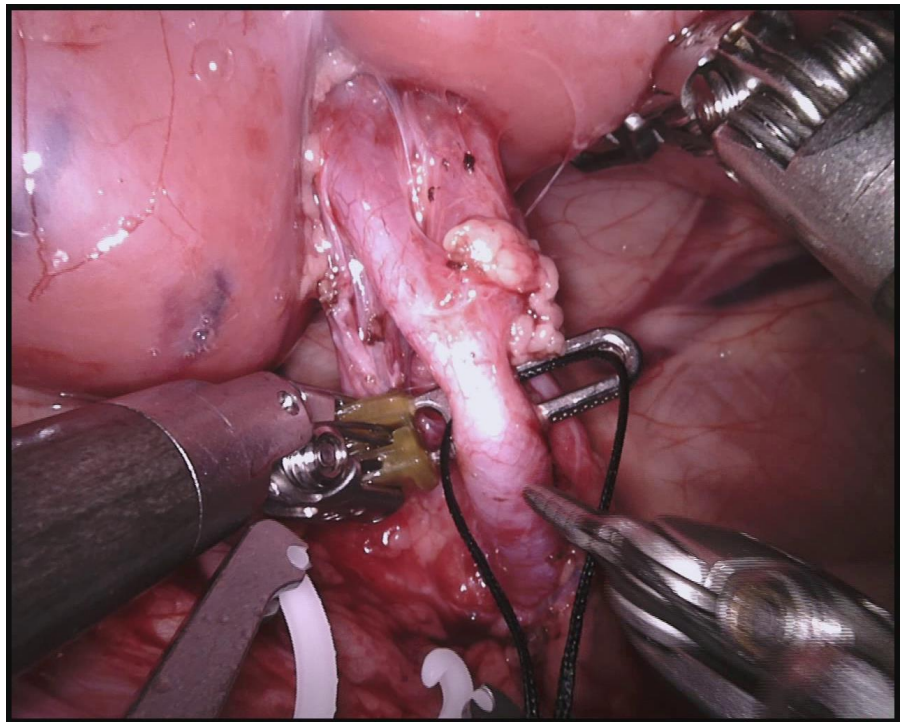
- Single keyword accuracy: 48.40%
- BLEU: 34.40%
- ROUGE-L: 53.64%
- METEOR: 52.69%
- LLM-as-a-judge Acc: 59.07%
- Cross Encoder Acc: 66.53%
- CLIP Text Similarity Acc: 82.67%

Truth:
In-template: 63.2%
Out-template: 27.07%
Total: 45.13%

Mentor: Dr. Laura Brattain

EndoVis-18-VQA Dataset

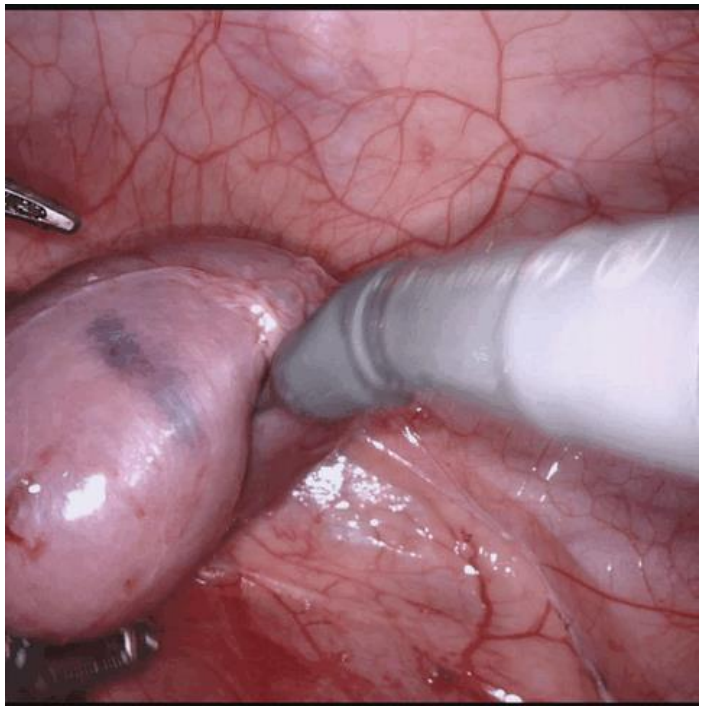
- Dataset contains 14 Nephrectomy videos
- 3 videos for testing
- Each video consists 148 frames
- Questions address -
  - anatomical structures,
  - tool–tissue interactions, and
  - instrument location
- Video clips made up of 8 frames
- QA pairs for each clip, 420 total
- In-template & out-of-template questions
- Long and short answers



Mentor: Dr. Laura Brattain

Example of EndoVis-18-VQA

Question: Where is prograsp forceps located?



Ground truth: prograsp forceps is located at left-top
Keyword: "left-top"

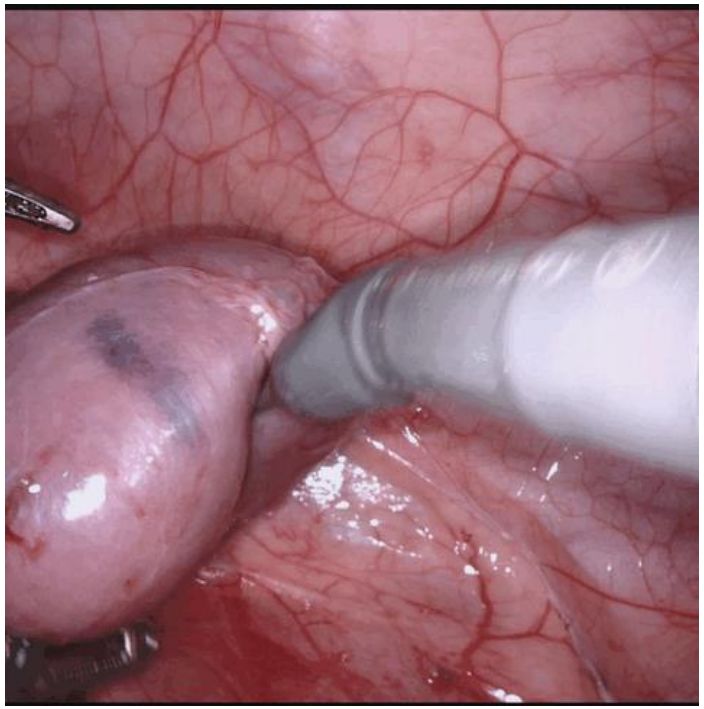
Prediction: prograsp forceps is located at left-top

Exact Match: True
Keyword Match: True

Mentor: Dr. Laura Brattain

Example of EndoVis-18-VQA

Question: What is the state of bipolar forceps?



Ground truth: action done by bipolar forceps is idle
Keyword: "Idle"

Prediction: action done by bipolar forceps is grasping

Exact Match: False
Keyword Match: False

Mentor: Dr. Laura Brattain

Current Results on Test Set (EndoVis-18)

- Single keyword accuracy: 48.77%
- BLEU: 78.96%
- ROUGE-L: 89.58%
- METEOR: 85.97%
- LLM-as-a-judge Acc: 78.73%
- Cross Encoder Acc: 42.13%
- CLIP Text Similarity Acc: 77.69%

Mentor: Dr. Laura Brattain

Next Steps

- Evaluate output
- Add a scene graph layer (e.g., PyTorch Geometric Temporal Libray)
- Expand to larger datasets

Mentor: Dr. Laura Brattain