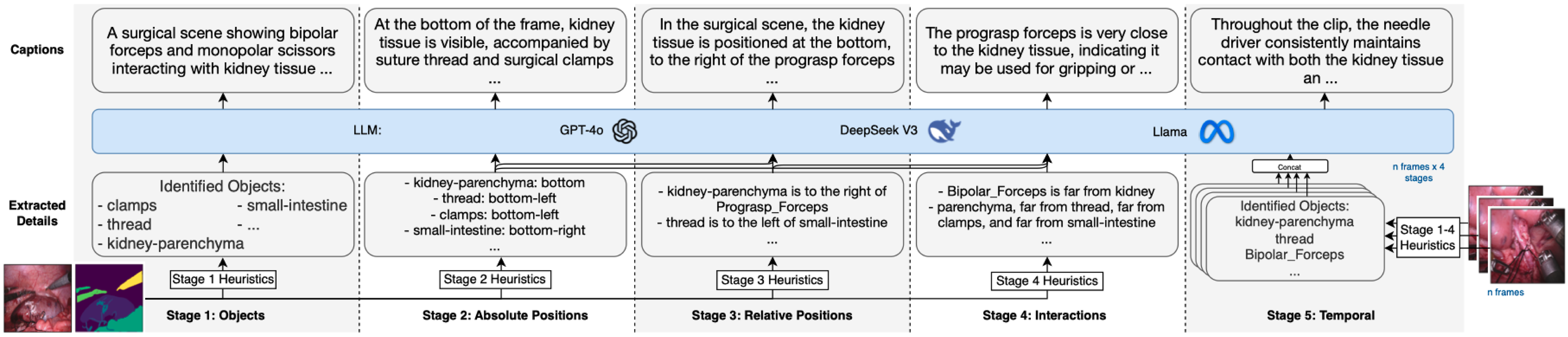
Segmentation-Informed Captioning: A Multi-Stage Pipeline for Surgical Vision–Language Dataset Generation

Mohamed Hamdy [a], Fatmaelzahraa Ali Ahmed [b], Mariam Ahmed [c], Mohannad AbuHaweeleh [c], Muraam Abdel-Ghani [b], Muhammed Arsalan [d], Abdulaziz Al-Ali [a], Shidin Balakrishnan [b]

[a] Computer Science and Engineering Department, College of Engineering, Qatar University, Doha, Qatar
[b] Department of Surgery, Hamad Medical Corporation, Doha, Qatar
[c] College of Medicine, Qatar University, Doha, Qatar
[d] KINDI Computing Research Center, College of Engineering, Qatar University, Doha, Qatar



Summary

Motivation:

- Surgical vision-language models (VLMs) require **high-quality** paired image-text data.
- Existing datasets (often based on **audio** transcriptions) are **noisy** and **poorly aligned**, limiting performance on **fine-grained** tasks like action recognition.

Core Contribution:

We propose a five-stage pipeline that generates descriptive and naturally sounding captions using existing segmentation datasets.

Pipeline Highlights:

- Extracts structured spatial and interaction cues in stages.
- Prompts large **language models (LLMs)** like GPT-4o to generate clean, natural captions.
- Avoids **error propagation** through **modular** stage-wise design.

Impact:

- Produces spatially and temporally grounded pseudo-captions.
- **95%** of generated captions rated **≥3** (out of 5) by medical experts.
- Enables better training data for **generalizable** surgical AI.

Results

Expert Evaluation:

- Medical experts rated captions across 5 stages from 3 LLMs: GPT-4o, Deepseek V3, LLaMA 3.3 70B.
- **95%** of captions scored **≥3**, and **73%** scored **≥4** on a 5-point Likert scale.

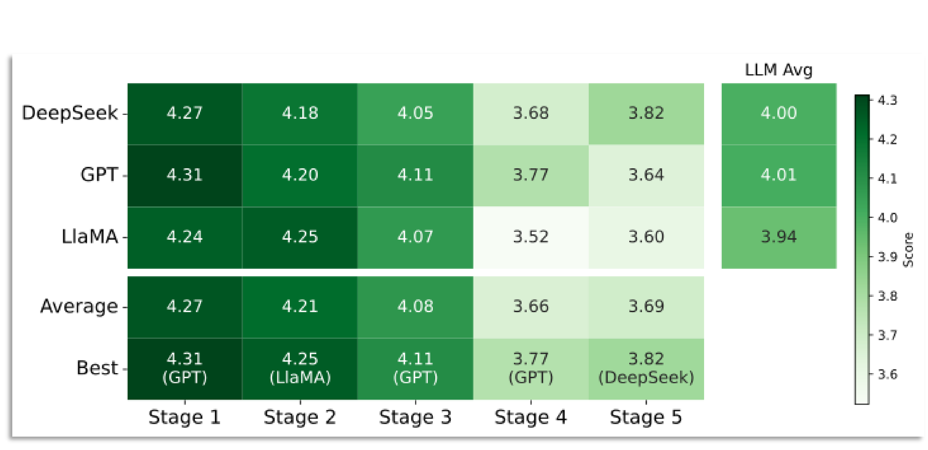
Stage-wise Trends:

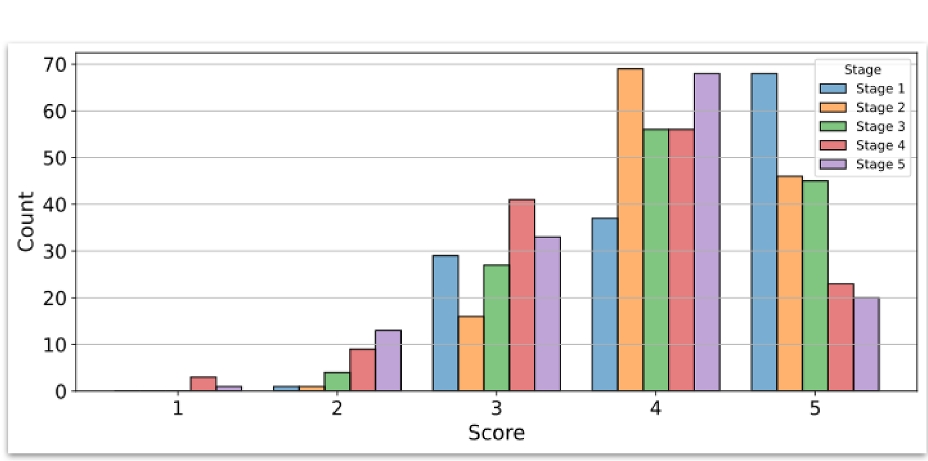
- **Highest scores:** Stage 1 (object listing) and Stage 2 (absolute positions).
- **Lowest scores:** Stage 4, due to ambiguity in proximity-based interaction inference.
- **Improvement in Stage 5** thanks to temporal context resolving ambiguities.

Model Comparison:

- **GPT-4o** consistently top-ranked (avg. rank: **1.97**) and never outperformed with statistical significance in any of the stages.
- **Deepseek V3** close second; **LLaMA 3.3 70B** performed worst in most stages.

	Stage 1		Stage 2		Stage 3		Stage 4		Stage 5		Overall	
	Rank	p-value	Rank	p-value	Rank	p-value	Rank	p-value	Rank	p-value	Rank	p-value
GPT-4o	**1.96**		2.00	0.564	1.98	0.914	**1.84**		2.07	0.169	**1.97**	
DeepSeek V3	2.00	0.527	2.04	0.527	2.11	0.874	1.98	0.509	**1.84**		2.00	0.867
LLaMA 3.3 70B	2.04‡	0.042	**1.96**		**1.91**		2.18‡	0.005	2.09‡	0.025	2.04	0.162





Methodology

Stage 1: Object Extraction

- **Objective:** Identify which **surgical instruments** and **anatomical structures** are visible in each frame.
- **Approach:** Extract labels directly from segmentation masks, without any spatial assumptions.
- **Outcome:** Produces **accurate but minimal** descriptions.

Stage 2: Absolute Positioning

- **Objective:** Add **absolute spatial context** to the detected objects.
- **Approach:** Divide the frame into regions (e.g., *top-left*, *center*) and assign object positions using overlap heuristics.
- **Outcome:** Captions become **anchored in the image space**, enabling location-aware prompts.

Stage 3: Relative Spatial Relationships

- **Objective:** Describe how objects are **positioned relative to one another.**
- **Approach:** Use mask dilation and centroid comparisons to infer pairwise relations like *"to the right of"* or *"on top of."*
- **Outcome:** Introduces **layout structure** into the scene, enhancing scene-level understanding.

Stage 4: Interaction Proximity

- **Objective:** Infer **how closely instruments interact** with anatomical targets — as a proxy for surgical actions.
- **Approach:** Simulate proximity using layered dilation and categorize interactions (e.g., *touching*, *very close*, *far*).
- **Outcome:** Adds **functional meaning** to captions, highlighting **potential** clinical intent.

Stage 5: Temporal Interaction Summary

- **Objective:** Capture **action over time** using multi-frame sequences.
- **Approach:** Aggregate spatial and interaction data across multi-frame clips to describe transitions like *"approaches"*, *"remains in contact", and actions like "grasping."*
- **Outcome:** Produces **video-level summaries** with temporal coherence — crucial for surgical training or analysis.

Prompting Large Language Models (LLMs)

- Each stage's structured data is turned into a **prompt** for a Large Language Model (LLM).
- Prompts are paired with a **stage-specific system message** that guides the tone, detail, and scope of the generated caption.
- Models like **GPT-4o**, **DeepSeek V3**, and **LLaMA 3.3 70B** are asked to produce **short**, **clinically coherent captions**.

Conclusion

High-Quality Surgical Captions from Segmentation Alone

- Our **five-stage pipeline** generates **clinically sound captions** by leveraging **spatial and temporal cues** from segmentation data, avoiding the noise and misalignment issues common in audio-based approaches.

Strong Expert Validation Across Stages

- **95% of captions** received scores ≥3, confirming **strong alignment** with stage-specific clinical expectations.

Foundation for Training Robust Surgical VLMs

- Provides a robust base for training **vision–language models** and enables future work in fine-tuning, benchmarking, and surgeon-led validation.

This work was supported by the Qatar Research Development and Innovation Council (QRDI) grant number ARG01-0522-230266.