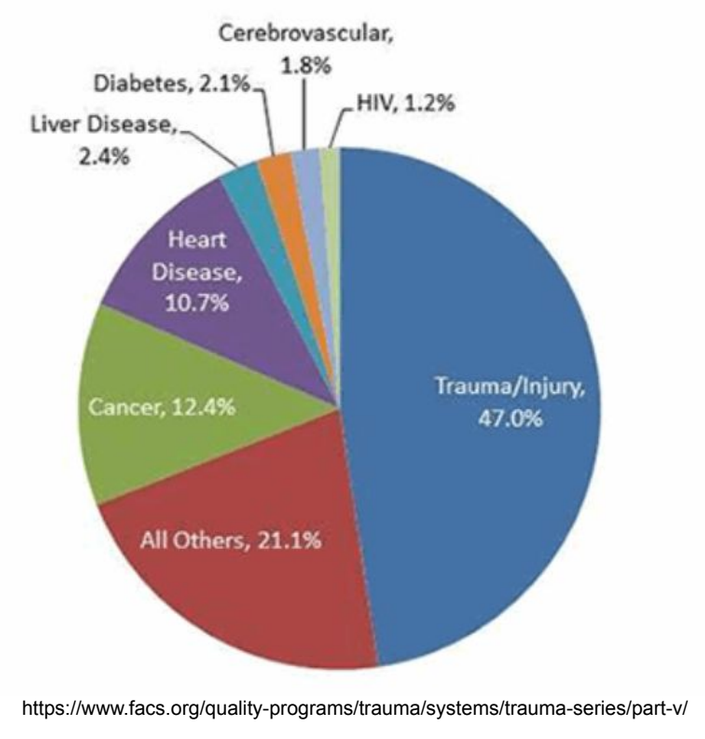
AI-Surgeon: A Transfer Learning Multimodal Ensemble Transformer Based VR Copilot for Action Prediction in Trauma Surgery

Issue Statement


https://allonehealth.com/how-to-cope-with-a-natural-disaster/

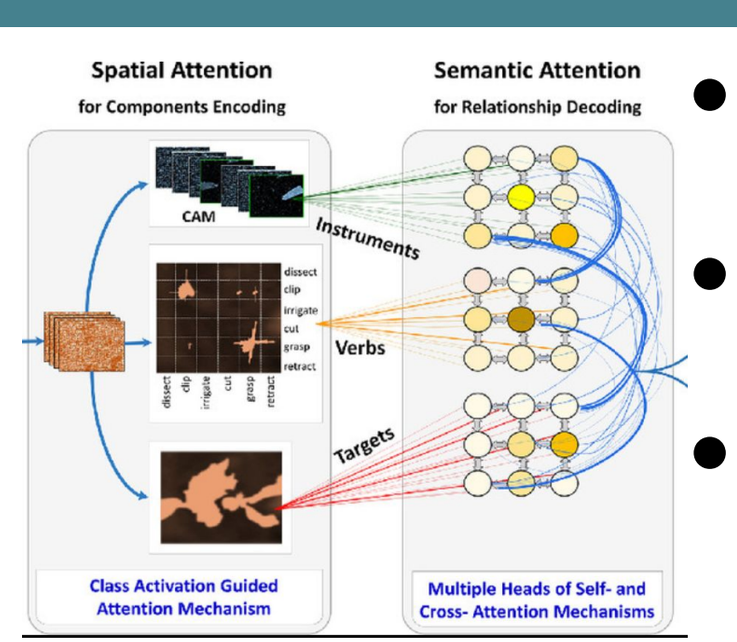

https://www.gettyimages.com/detail/news-photo/firefighters-escape-from-a-collapsed-building-during-a-news-photo/1904688449?adppopup=true

- **Life-saving surgical procedures** for trauma injuries require immediate actions, but in **austere situations**, such as disaster zones or remote areas, there are often no immediate medical professionals available
- First responders often **lack the necessary medical skills** and resources to perform life saving surgery accurately and quickly
- **Six million deaths** around the world are caused by trauma injuries annually, with **60% occurring just minutes after the injury**
- Current **machine learning and AI assistants** in trauma surgery are still In their early development stages



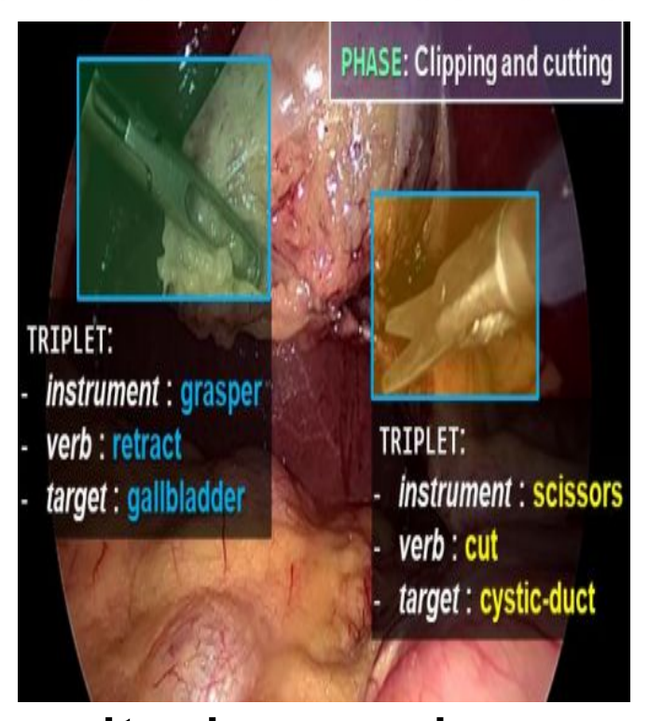
https://www.facs.org/quality-programs/trauma/systems/trauma-series/part-v/

Related Works in Surgical Aid



- Architecture framework improvements for surgical guidance [1]
- Mainly focusing on Coarse-Grained phase recognition currently
- Lack of tool-based interactions and detection for fine-grain real time surgical action recognition and anticipation
- Parameter efficient transformer methods

- Curation of larger and comprehensive surgical datasets is needed [2]
- Multi-view surgical action datasets from different point of views of different people in the operating room
- Action triplet/duplet development and curation in laparoscopic surgery



- Development of telemonitoring and virtual reality (VR) systems [3]
- Use of VR in training of surgical procedures
- Telemonitoring and remote assistance in operating rooms applications
- VR for improving patient rehabilitation

Research Gaps and My Solutions

Research Gap	Solution
Most trauma injuries occur in austere and difficult to access environments	Development of an offline VR HMD system to provide real-time assistance to first responders
First responders lack equipment and training necessary to perform accurate, efficient LSI procedures	Leverage AI and machine learning to guide first responders through the procedure
Most existing works recognize surgical activities at a coarse-grained level, such as phases, steps or events	Recognizing surgical actions as <verb, target > action pair delivers more comprehensive details
Current ML frameworks for action recognition and action anticipation low in accuracy and computationally expensive	Utilize spatiotemporal relationships of surgical procedures and transfer learning to improve accuracy and reduce computational complexity
Lack of comprehensive and large scale datasets for trauma surgery	Curate the Trauma THOMPSON Dataset, the first comprehensive egocentric medical dataset

Research Objectives

- Develop an **offline** virtual reality head mounted device (HMD) as a **copilot** to assist medical professionals, personnel with limited training, and first responders in trauma surgery in resource-constrained and remote environments
- Curate **egocentric video dataset** for trauma surgery procedures with <verb, target> pair action annotations
- Implement **spatiotemporal** AI frameworks for real-time surgical action recognition and anticipation
- **AI-Surgeon is about leveraging AI research to save lives and increase access to medical care**

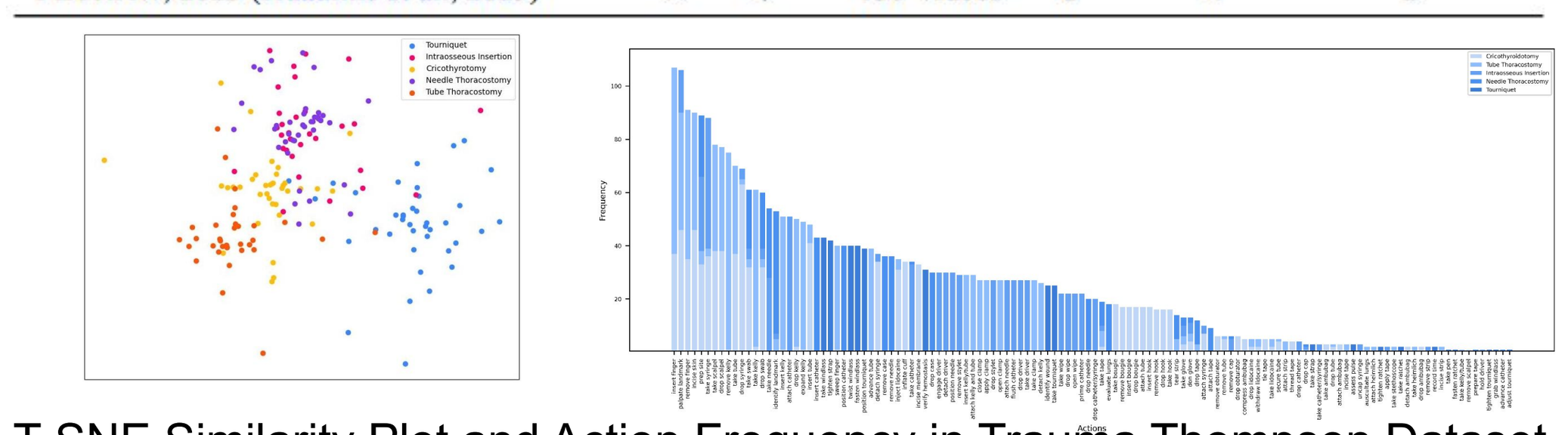
Design Criteria and Constraints

Criteria	Constraints
Action recognition models must be able to predict <verb-, target> pairs	Algorithms implemented on HMDs must be lightweight
Action recognition accuracy of minimum 50% and anticipation accuracy of minimum 40%,Visual Question answering accuracy of minimum 70%	Surgeries are commonly performed in resource-limited and rural environments
Lightweight and offline processing for deployment in VR for use in austere environments	Difficulty of collecting high quality data for surgical procedures

Egocentric Trauma Surgery Video Dataset

Overview and Sample Frames including Verb-Target Pairs From Trauma Thompson Dataset

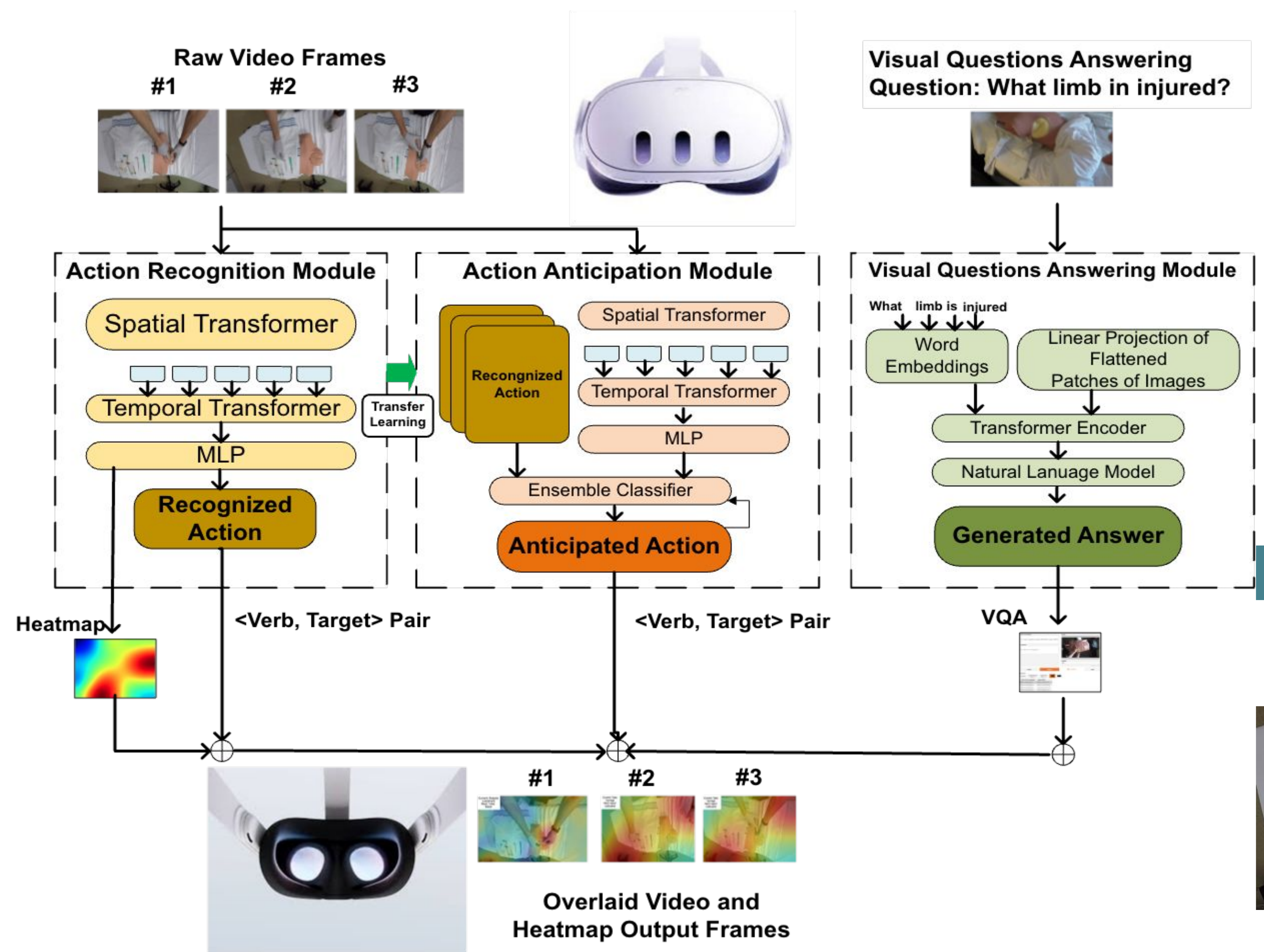
Dataset	Ego	Med	Frames	No. Act	Participants	No. Envs
Trauma THOMPSON, 2025	✓	✓	0.7M	162	12	15
EPIC-Kitchens, 2018 (Damen et al., 2018)	✓	×	11.5M	149	32	32
BEOID, 2014 (Damen, 2014)	✓	×	0.1M	34	5	1
GTEA, 2011 (Fathi et al., 2011)	✓	×	0.4M	42	13	1
CMU-MMAC, 2008 (de la Torre et al., 2008)	✓	×	0.2M	31	16	1
ADL, 2012 (Pirsiavash & Ramanan, 2012)	✓	×	1.0M	32	20	20
ESAD, 2020 (Bawa et al., 2020)	×	✓	0.03M	21	4	4
CholecT50, 2022 (Nwoye et al., 2022)	×	✓	0.1M	100	13	13
MedVidCL, 2023 (Gupta et al., 2023)	×	✓	1489 Videos	0	>100	>100
MRAO, 2021 (Schmidt et al., 2021)	×	✓	480 Videos	10	16	2
MISAW, 2021 (Huaulmé et al., 2021)	×	✓	27 Videos	17	6	1
PSI-AVA, 2022 (Valderrama et al., 2022)	×	✓	8 Videos	167	3	1
PETRAW, 2023 (Huaulmé et al., 2023)	×	✓	150 Videos	6	4	2



T-SNE Similarity Plot and Action Frequency in Trauma Thompson Dataset

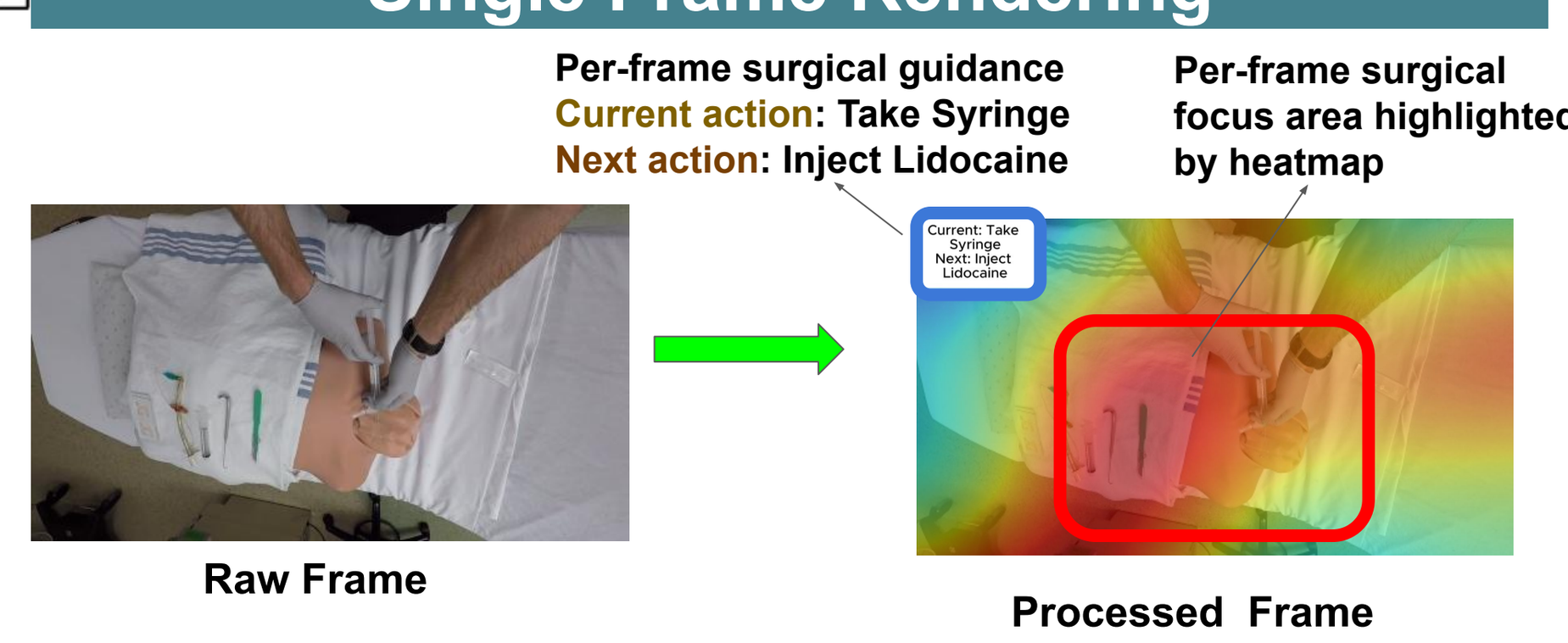
- **The Trauma Thompson dataset** offers **5 trauma procedures** (left picture), **162 actions**, and **700,000 video frames** and is the only **egocentric** medical dataset **annotated by medical professionals** currently available for **trauma surgery**

AI-Surgeon Overall Architecture



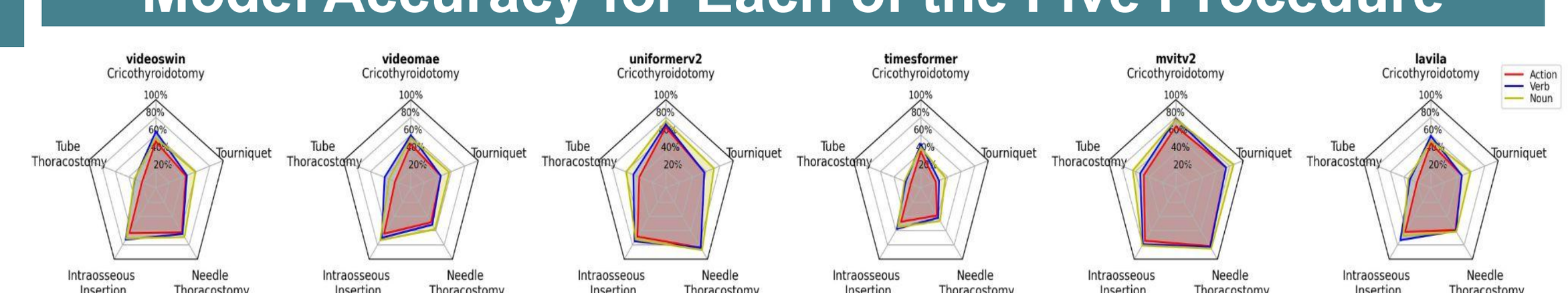
- **Module 1, Action Recognition:** MViTv2 (Multiscale Vision Transformer) deciphers surgeon's current actions into a **<verb, target> action pair**, with a **Grad-CAM** heatmap highlighting the area of focus
- **Module 2, Action Anticipation:** Predicts the next surgical action using novel **Surgiformer** network, which utilizes **transfer learning** and temporally adaptive **ensemble** transformer to memorize previous video frames and capture the **sequential** nature of workflows, which current deep learning models struggle with.
- **Module 3, Visual Question Answering (VQA):** Provides real-time feedback to surgery-related questions using an image-text transformer decoder and a generative AI encoder

Single Frame Rendering



Module 1: Action Recognition

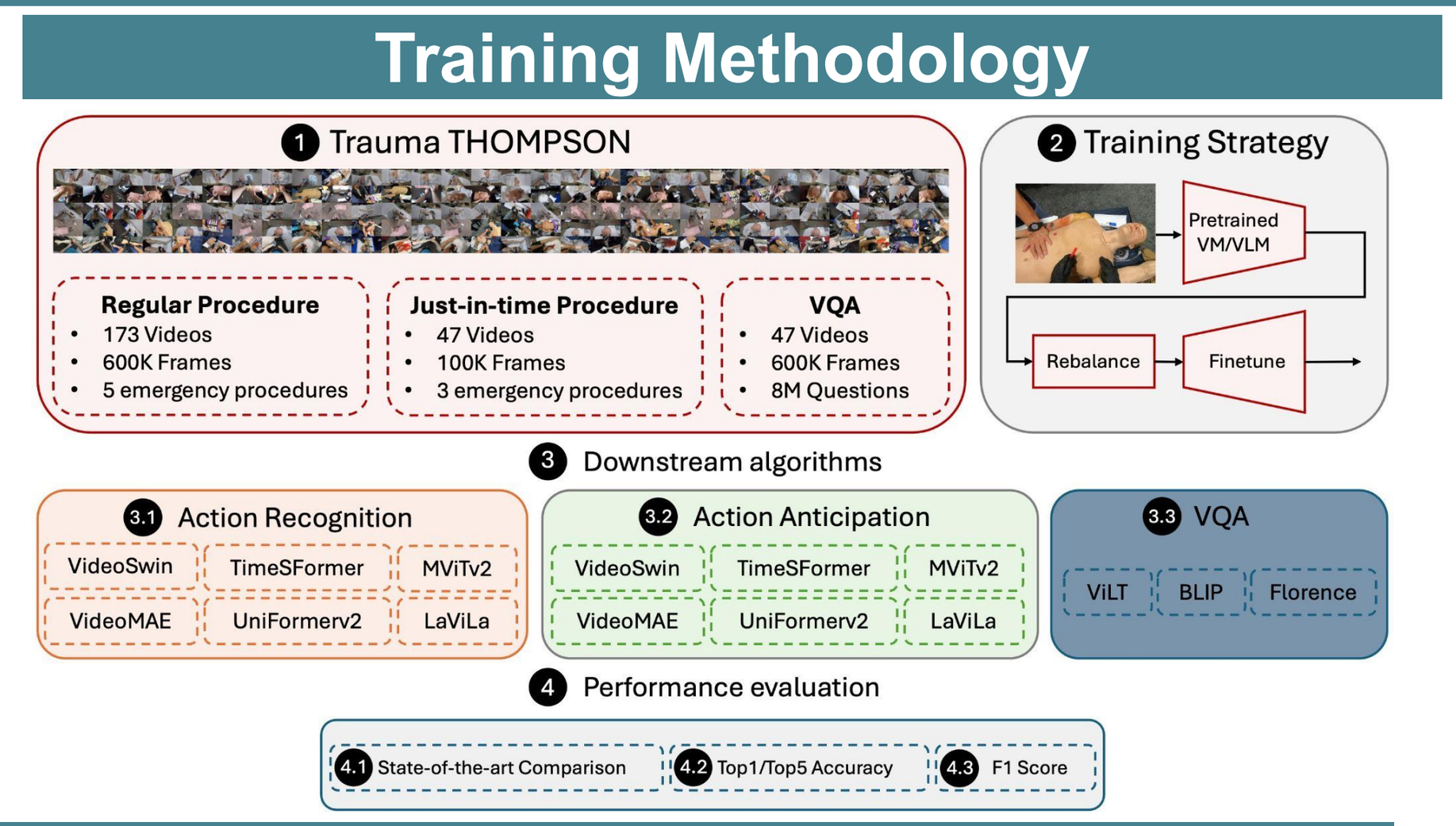
Model Accuracy for Each of the Five Procedure



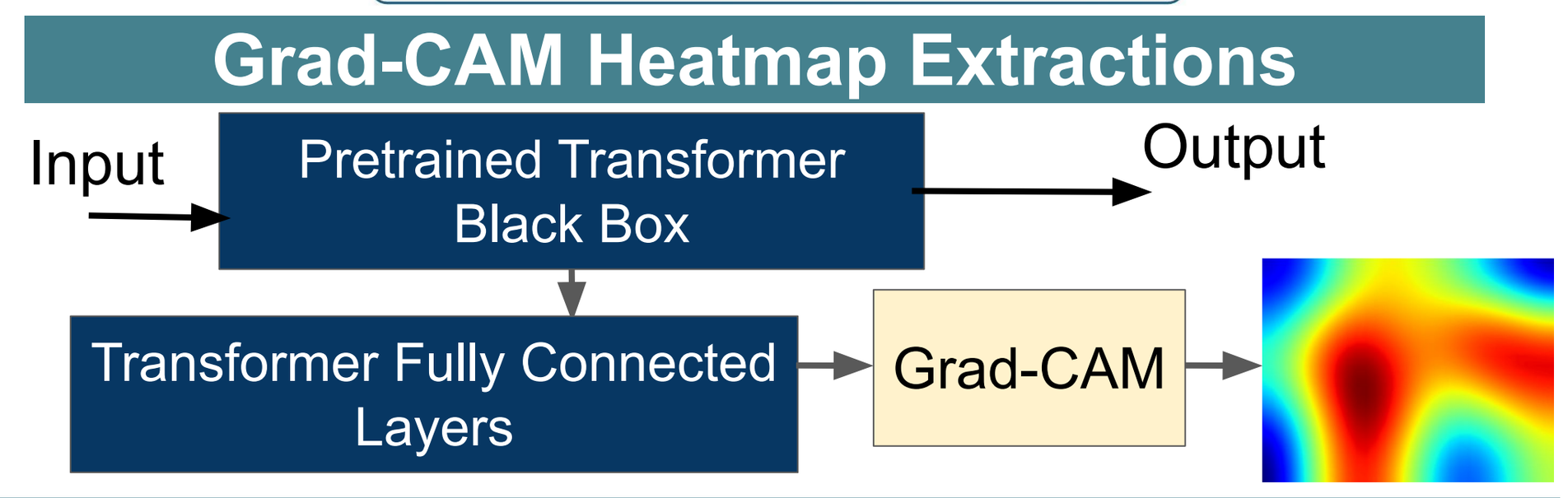
Model Accuracy of Different Transformers

Model	Dataset	Train Frame Rate	Validation Frame Rate	Data Augmentation Method	Top-1 Accuracy
MVITv2	Regular	32x2x1	32x2x4	CenterCrop	**65.59**
VideoSwin	Regular	32x2x1	32x2x4	CenterCrop	45.10
UniformerV2	Regular	32x2x1	32x2x4	CenterCrop	60.47
VideoMAE	Regular	32x2x1	32x2x4	CenterCrop	43.34
TimeSFormer	Regular	32x2x1	32x2x4	CenterCrop	31.91
LaVila	Regular	16x2x1	16x2x1	CenterCrop, TemporalCrop	42.52
Previous Works	Regular	-----	-----	-----	**45.10**

Training Methodology

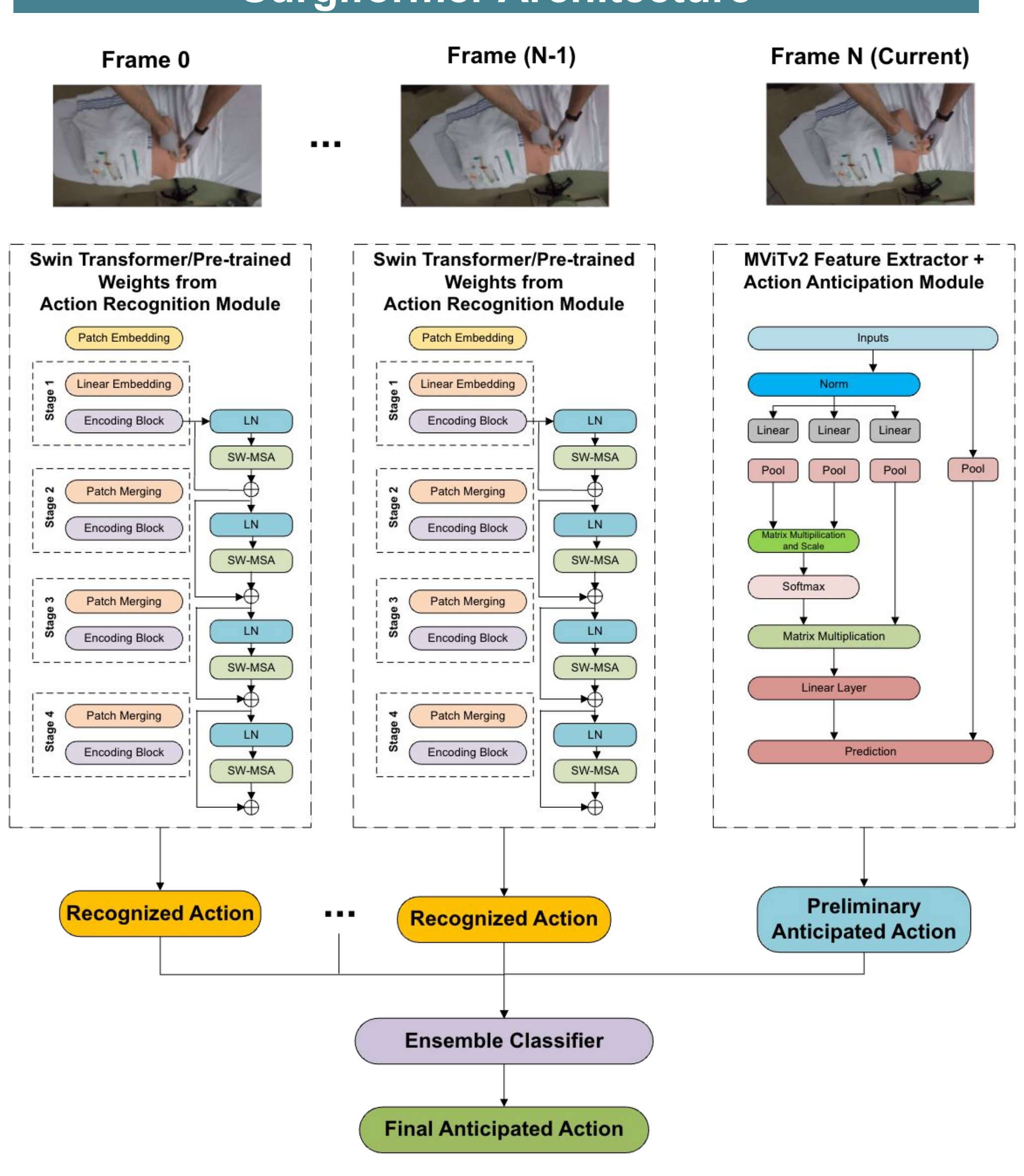


Grad-CAM Heatmap Extractions

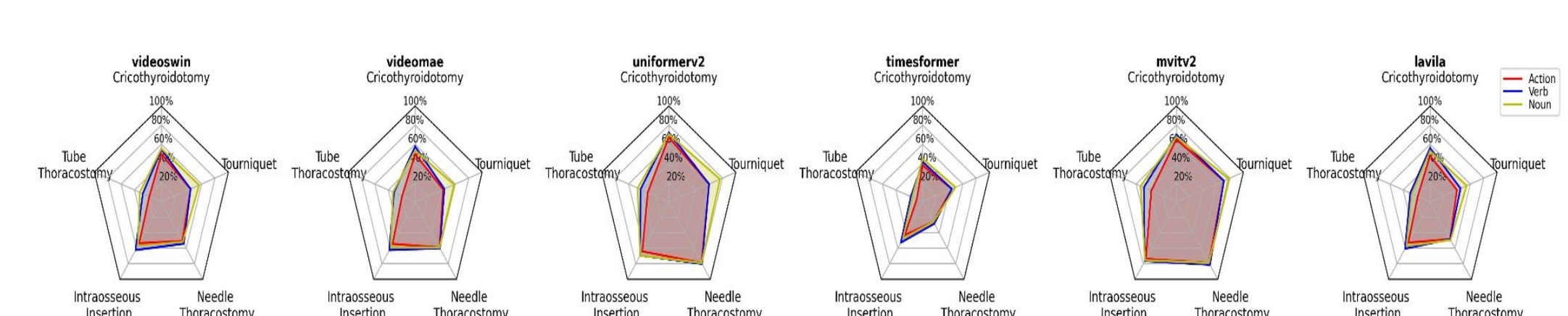


Module 2: Action Anticipation with Surgiformer

Surgiformer Architecture



Individual Model Accuracy for Each Procedure



Ablation Studies for Surgiformer Network

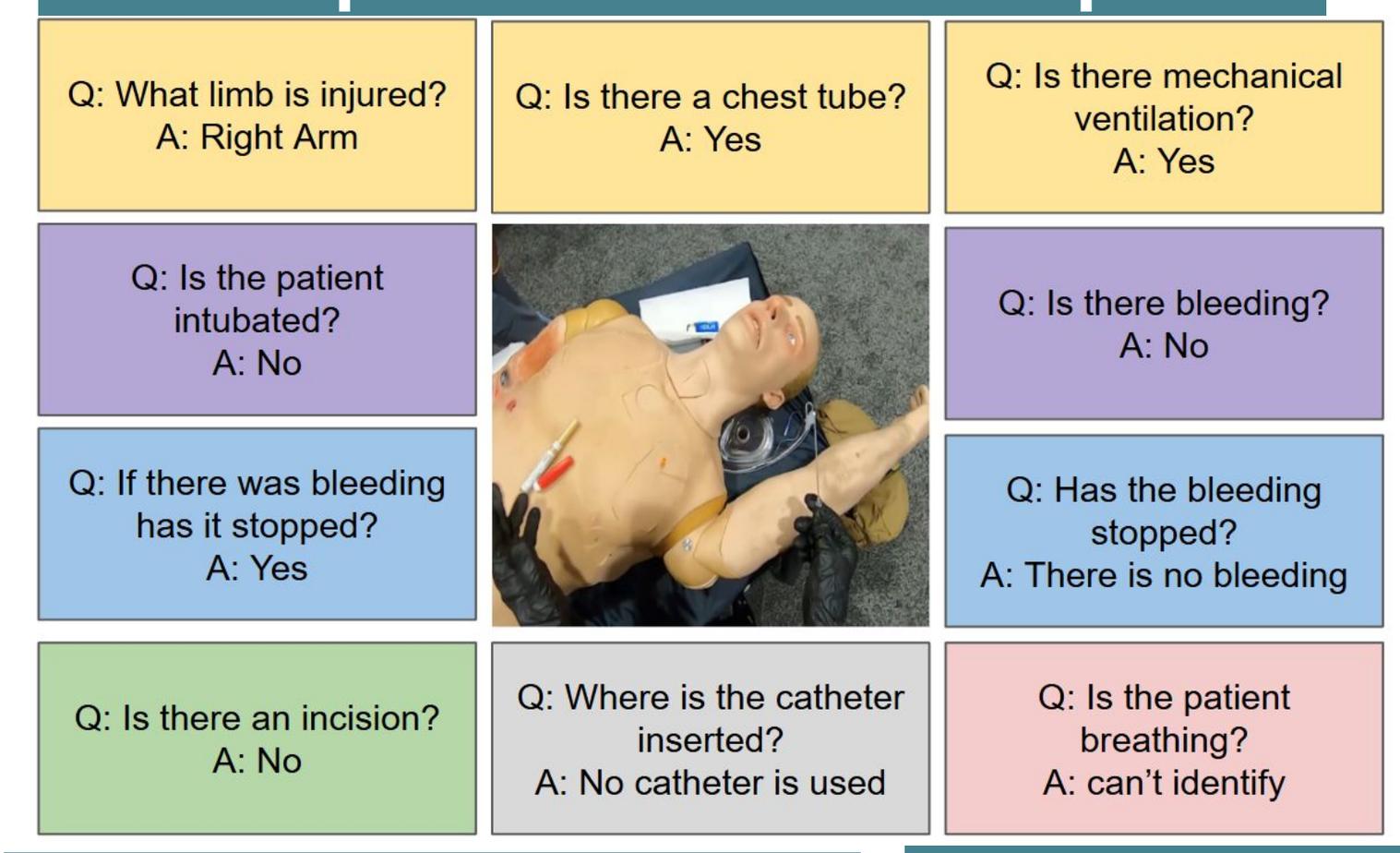
Model Combination (Recognition, Anticipation)	Accuracy	Model Combination (Recognition, Anticipation)	Accuracy
VideoSwin, MViTv2	**81.81**	VideoSwin, VideoMAE	73.41
UniformerV2, MViTv2	81.60	MViTv2, VideoMAE	64.61
MViTv2, MViTv2	77.74	UniformerV2, VideoMAE	60.59
VideoMAE, MViTv2	65.69	VideoMAE, VideoMAE	53.32
VideoSwin, UniformerV2	80.37	VideoSwin, VideoSwin	74.81
VideoMAE, UniformerV2	78.67	UniformerV2, VideoSwin	65.69
UniformerV2, UniformerV2	71.41	MViTv2, VideoSwin	64.61
MVITv2, UnifrormerV2	70.79	VideoMAE, VideoSwin	51.93

Selection Process for Surgiformer Architecture

- Selection of transformer models based on accuracy
- Combinations of top 4 most accurate models in action recognition tested (VideoSwin, VideoMAE, UniformerV2, MViTv2), used most accurate combination in final Surgiformer model
- Previous works achieve around 40%

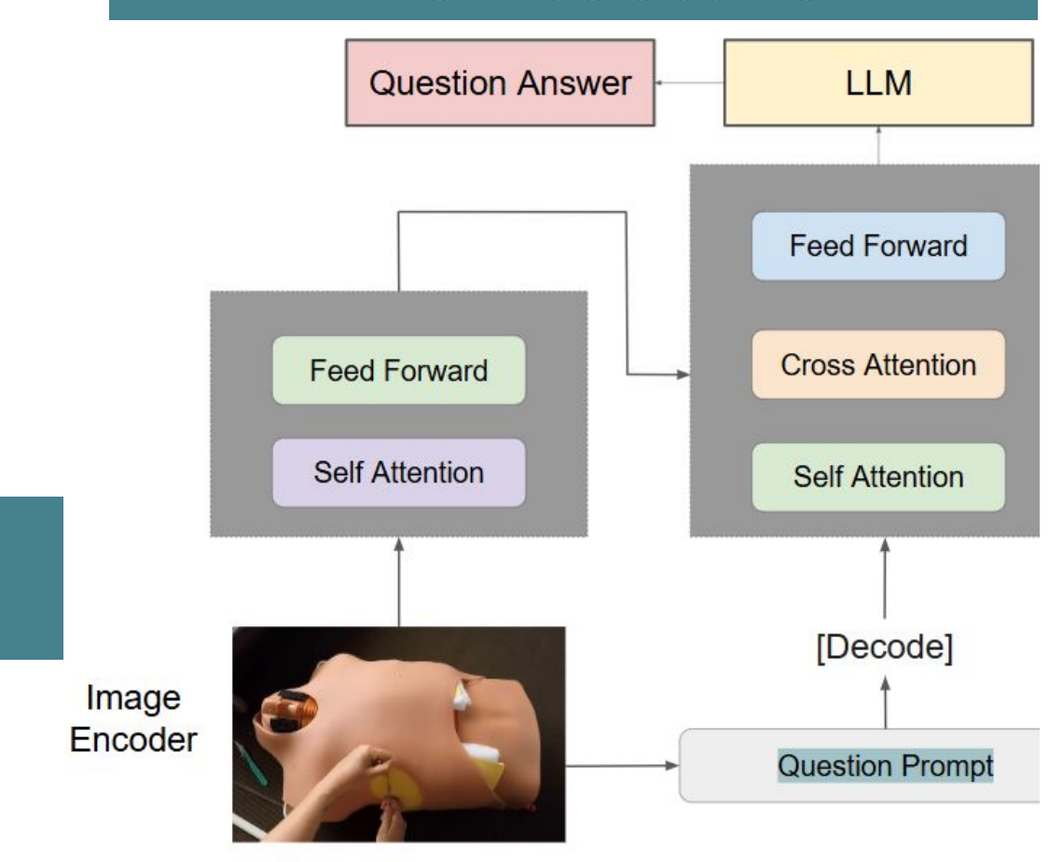
Module 3: Visual Question Answering (VQA)

Prompt and Answer Samples



- Dataset Annotations based on a given image
- Over 8 Million questions in total
- Multimodal text and image inputs

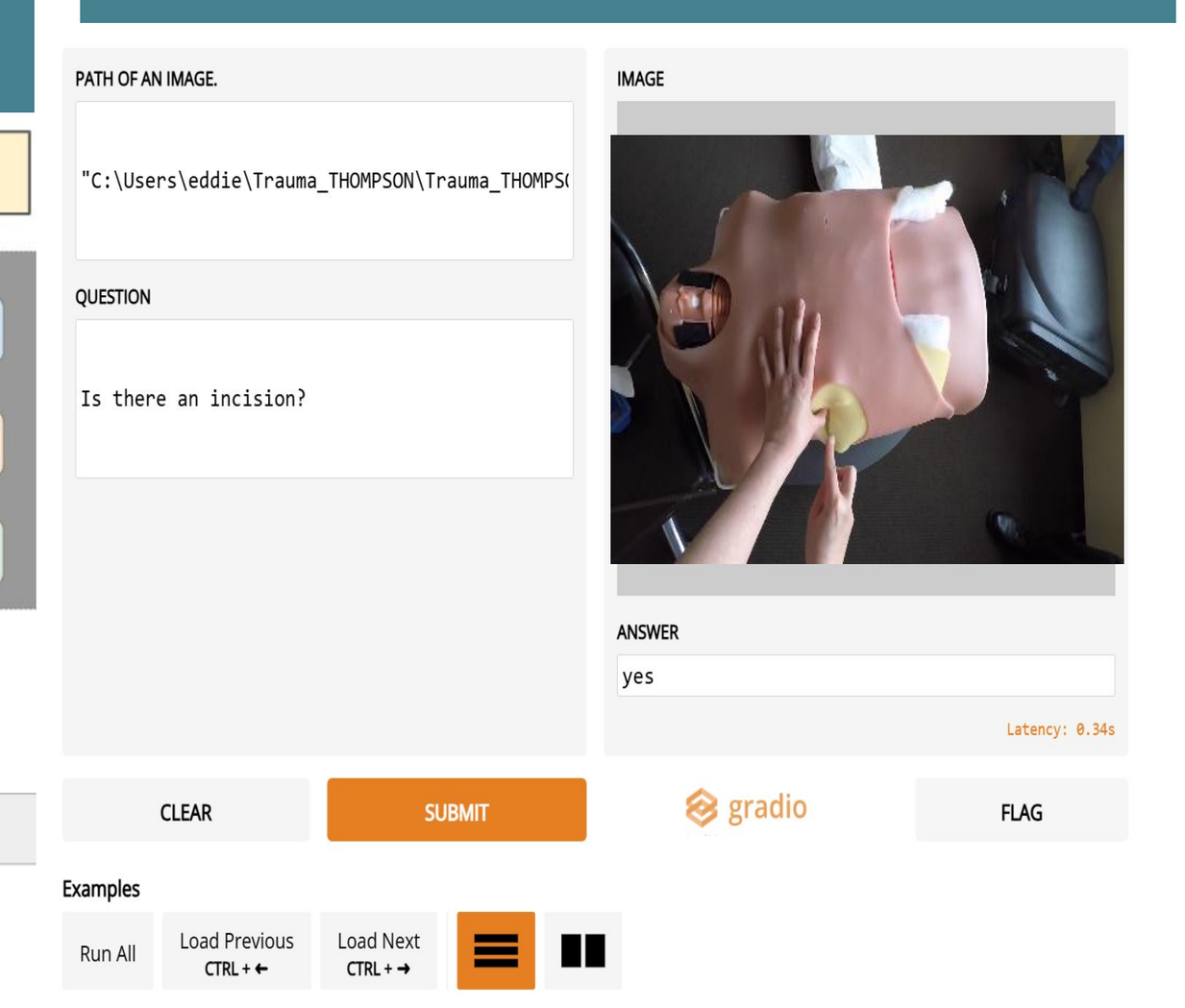
BLIP Model Architecture



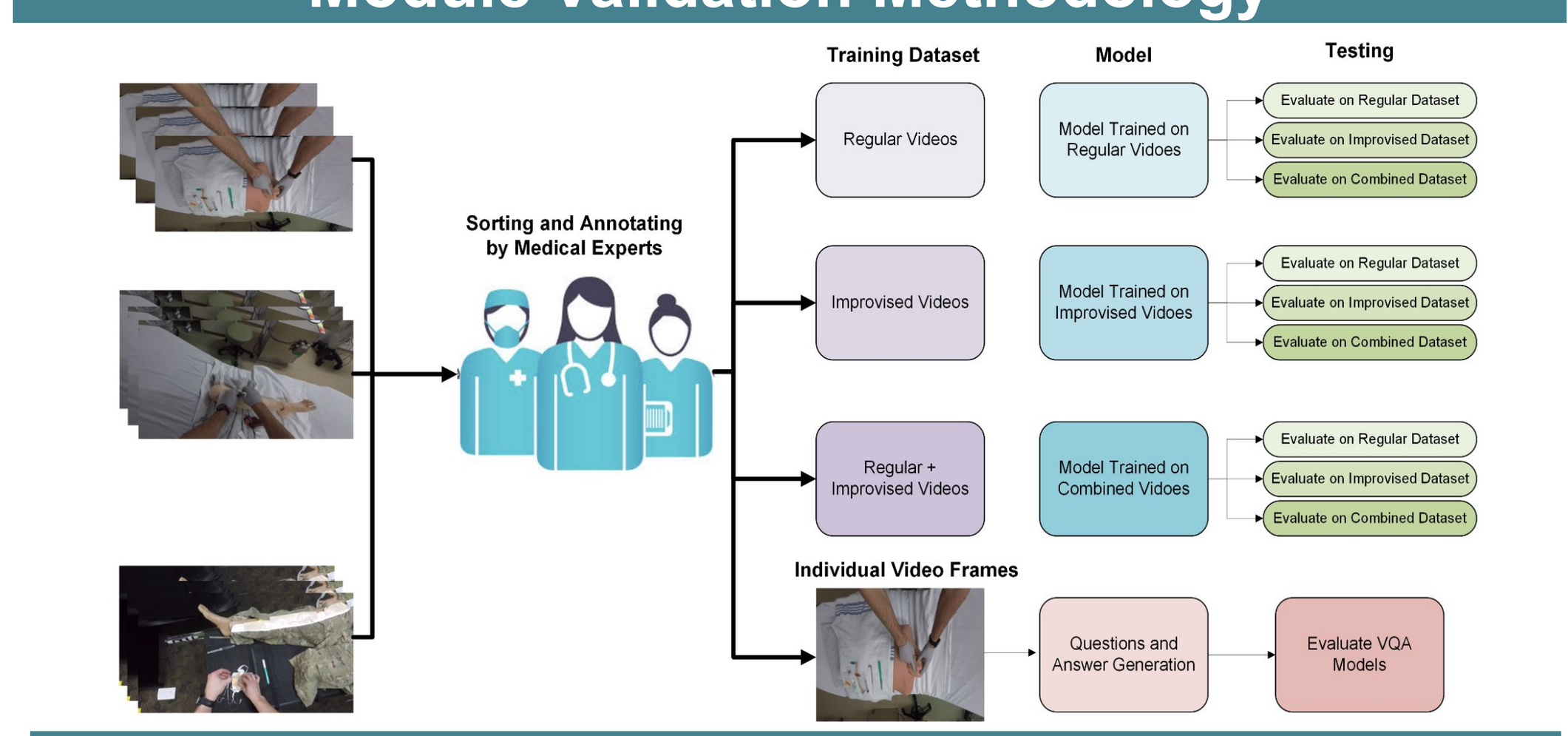
BLIP Model Achieved Highest Accuracy

Model	Accuracy	F1Score
ViLT	79.88	80.00
BLIP	88.64	100.00
Florence	87.86	100.00

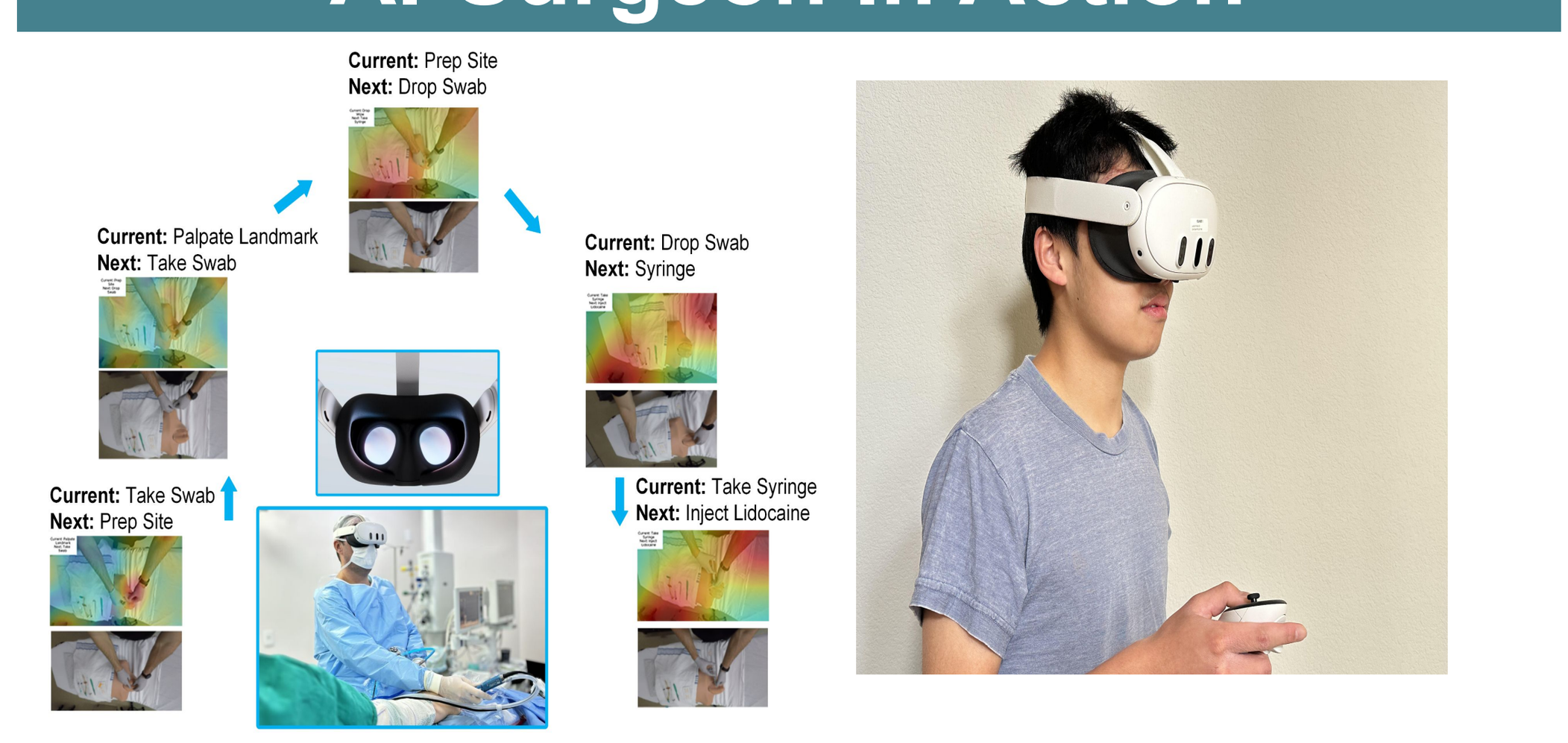
VQA User Interface

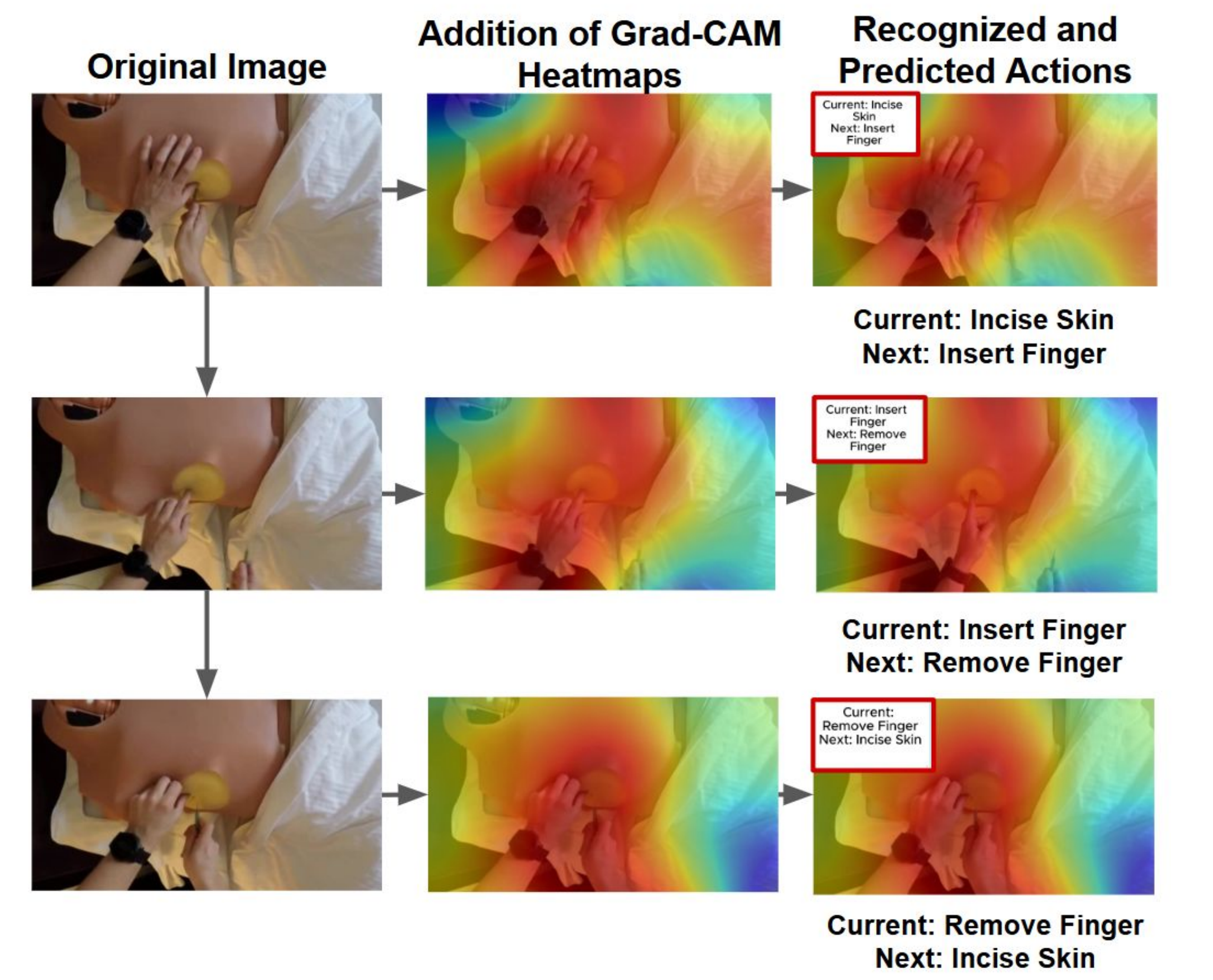


Module Validation Methodology



AI-Surgeon In Action





Conclusions and Future Works

- **AI-Surgeon**, an AI-based surgical VR copilot is implemented to provide **offline intraoperative guidance** for medical procedures in austere, uncontrolled and low medical resource settings
- The **Surgiformer** network deploys a temporally adaptive memory augmented ensemble transformer and transfer learning to achieve high accuracy for action anticipation
- On the **Trauma Thompson dataset**, the largest egocentric medical dataset, AI-surgeon achieves 65.59%/81.81%/88.64% in accuracy for action recognition, action anticipation, and VQA respectively, significantly higher than existing methods
- AI-Surgeon can also be deployed for **hospital workflow optimization**, performance evaluation and surgical training etc.
- **Future Works**: Implementing more assistance modules (hand tracking, tool detection)

Selected References

1. Nwoye, C. I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., & Padoy, N. (2022). Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis, 78*, 102433. https://doi.org/10.1016/j.media.2022.102433
2. Ríos, M. S., Molina-Rodriguez, M. A., Londoño, D., Guillén, C. A., Sierra, S., Zapata, F., & Giraldo, L. F. (2023). Cholec80-CVS: An open dataset with an evaluation of Strasberg's critical view of safety for AI. *Scientific Data, 10*(1), 194. https://doi.org/10.1038/s41597-023-02073-7
3. McKnight, R. R., Pean, C. A., Buck, J. S., Hwang, J. S., Hsu, J. R., & Pierrie, S. N. (2020). Virtual Reality and Augmented Reality-Translating Surgical Training into Surgical Technique. *Current Reviews in Musculoskeletal Medicine, 13*(6), 663–674. https://doi.org/10.1007/s12178-020-09667-3
4. Schreiter, J., Heinrich, F., Hatscher, B., Schott, D., & Hansen, C. (2024). Multimodal human–computer interaction in interventional radiology and surgery: A systematic literature review. *International Journal of Computer Assisted Radiology and Surgery*. https://doi.org/10.1007/s11548-024-03263-3