Context-Aware Vision Language Model for Action Recognition

Authors: Eddie Zhang[1], Yupeng Zhuo[2], Juan Wachs[2*]

Affiliations: 1: The Harker School, USA; 2: Purdue University, USA

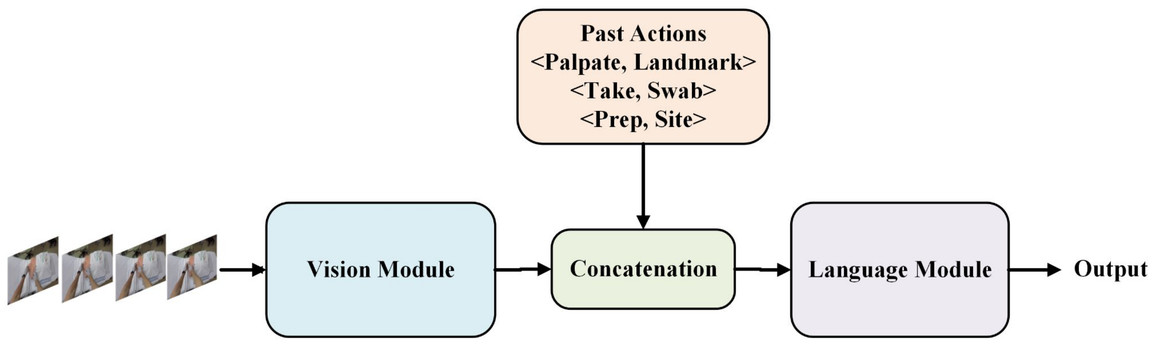
Research Goals

- Demonstrate the feasibility of **vision language models** on surgical **action recognition** and **action anticipation**
- Present a **memory-augmented** VLM that improves its **contextual understanding** of the current procedure through **memorizing** and **concatenating** the **current** recognized **action** with **past actions**
- Evaluate the model on the challenging **Trauma THOMPSON dataset** for action recognition and action anticipation demonstrating the **superior performance** of the proposed architecture over results reported by the previous literature

Introduction

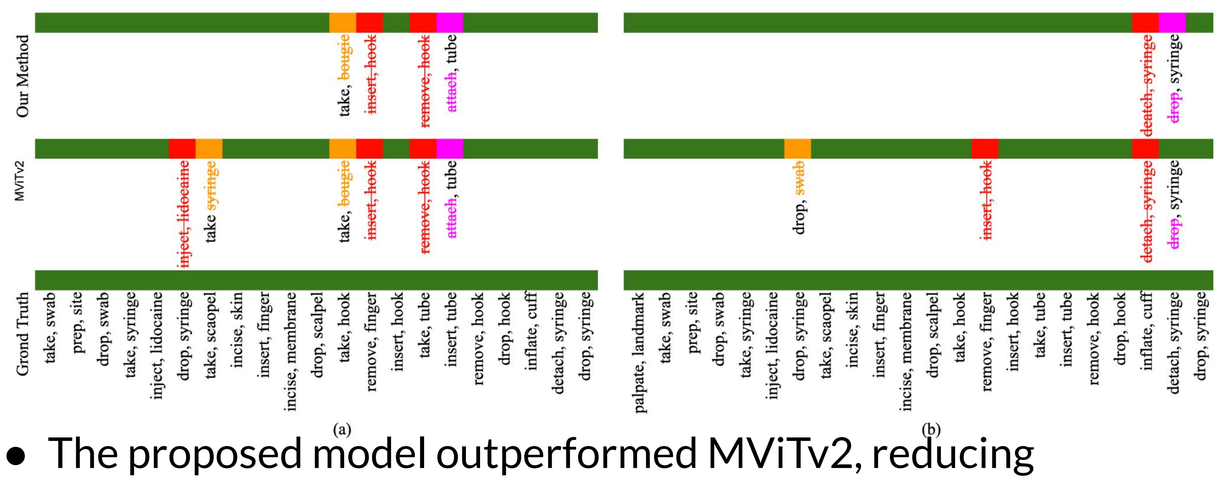
- **Humanitarian medicine** delivers care in crises like war and disasters, often in remote areas with rising patient complexity and limited provider capacity. **Telemedicine** (e.g., telementoring) helps guide responders but struggles in disconnected or conflict zones.
- **AI copilots** assist in surgery by enhancing precision and outcomes through real-time support, using **action recognition** to interpret movements, improve coordination, and enable automated documentation.
- **Anticipation** features let AI predict future surgical steps, suggest tools, adjust imaging, and warn of complications.
- Combined, **action recognition and anticipation** reduce surgeon cognitive load, boost efficiency, and improve surgical performance.

Methodology



- **Surgical actions** are defined as **<verb, noun>** pairs (e.g., <take, swab>), enabling structured recognition and prediction of procedural steps.
- The proposed model combines a **vision module** for extracting visual features and predicting actions from video segments, and a **language module** that uses past actions to learn patterns and anticipate future ones.
- The **vision module** consists of the MViTv2 Transformer to extract video frame features
- The **language module** uses memory of all past predicted actions in ap procedure by processing concatenated past and current predictions through the FlanT5 LLM.
- By leveraging the **sequential nature of surgical actions**, the language model improves temporal context and accuracy, using input features $X_i = \text{concat}(x_0, x_1, x_2, \ldots, x_n)$ to predict a sequence of actions $y = (y_0, y_1, y_2, \ldots, y_i)$

Results



- The proposed model outperformed MViTv2, reducing recognition errors from **9 to 6** (a **33% improvement**) and anticipation errors from **6 to 3** (a **50% improvement**).
- These gains come from the model's ability to **learn contextual patterns** in surgical action sequences, **combine vision and language outputs**, and even **correct earlier prediction errors** as the procedure progresses.

Model	Top1	Top5	FLOPs	Params
LaViLA	42.52	68.81	70T	3.48B
Timesformer	31.91	62.81	196G	122M
VideoSwin	45.10	74.52	180G	87M
VideoMAE	43.34	71.89	88G	28.2M
UniformerV2	60.47	85.65	100G	115M
MViTv2	65.59	**89.75**	64G	34.5M
MViTv2+FlanT5	**69.11**	87.26	76G	111.5M

Model	Top1	Top5	FLOPs	Params
LaViLA	38.79	67.85	70T	3.48B
Timesformer	28.44	59.97	196G	122M
VideoSwin	39.88	69.24	180G	87M
VideoMAE	41.42	69.24	88G	28.2M
UniformerV2	56.25	84.70	100G	115M
MViTv2	60.12	**87.02**	64G	34.5M
T5 + MViTv2	**62.13**	81.14	76G	95M

- The **proposed method** is able to achieve an **accuracy** of **69.11%** on **action recognition** and **62.13%** on **action anticipation**, compared to **65.59%** for action recognition and **60.12%** for action anticipation for the **MViTv2 model**

Discussion & Conclusion

- The proposed method outperforms vision-only models on longer procedures, with **low latency** (2.65s)

ViT	Recognition	Anticipation	FLOPs	Params
VideoMAE	50.36	43.43	180G	87M
VideoSwin	51.10	44.67	88G	28.2M
UniformerV2	31.77	40.80	100G	115M
MViTv2	**69.11**	**62.13**	64G	34.5M

LLM	Recognition	Anticipation	FLOPs	Params
BART	67.20	53.48	25G	139M
BERT	54.32	57.65	56G	247M
T5	66.03	**62.13**	129G	569M
PEGASUS	37.71	27.98	12G	60.5M
Flan-T5	**69.11**	61.21	12G	77M

Ablation Studies

- **Future work** includes extending the model to **improvised or "just-in-time" procedures** and exploring a **broader range of surgical tasks**.

The Trauma THOMPSON Challenge 2025 Signup