The Trauma THOMPSON Dataset for Real-World Emergency AI

Yupeng Zhuo[1], Eddie Zhang[1], Xiangchen Yu[1], Aditya Pachpande[1], Andrew W. Kirkpatrick[2], Kyle Couperus[3], Jessica Mckee[2], Juan Wachs[1*]

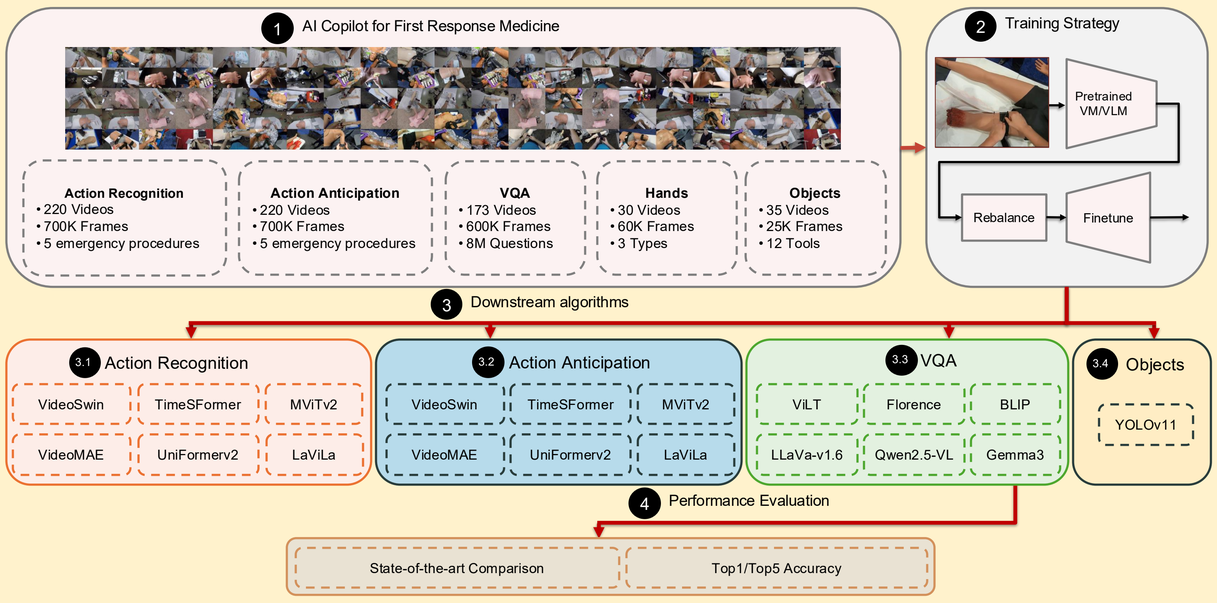
[1]Purdue University, West Lafayette, IN, USA; [2]University of Calgary, Calgary, Alberta, Canada; [3]The Geneva Foundation, Tacoma, WA, USA
*Email: jpwachs@purdue.edu

Motivation

- Providing high-quality medical care in resource-limited settings (e.g., disaster zones, rural areas) is challenged by a lack of expertise, limited equipment, and poor connectivity.
- First responders often have minimal training yet must handle complex medical cases, increasing the risk of poor patient outcomes.
- There is a critical need for autonomous, AI-driven systems that can provide real-time guidance and decision support to medical personnel in high-risk, disconnected environments.
- This work introduces an AI-powered copilot system to assist with life-saving interventions, trained on a novel dataset designed for humanitarian operational medicine.

Challenge: Insufficient Data Availability

Trauma THOMPSON



Key Highlights

- **220** videos
- **5** critical life-saving procedures
  - Cricothyroidotomy
  - Intraosseous Infusion
  - Needle Thoracostomy
  - Tube Thoracostomy
  - Tourniquet
- Indoor + outdoor, natural mistakes, blood, occlusions.
- Combines regular procedures and just-in-time (JIT) improvised procedures.
- Dataset size
  - **0.7M** frames
  - **3717** clips
  - **124** actions for regular procedures
  - **86** actions for JIT procedures
- Multi-level annotations: actions, hands, objects, VQA.
- Benchmarks in action recognition, anticipation, VQA.
- Open access
  DOI: 0.7910/DVN/V5BTRU

TRAUMA THOMPSON
CLINICAL DECISION SUPPORT FOR THE FRONTLINE MEDIC

Gemma 3

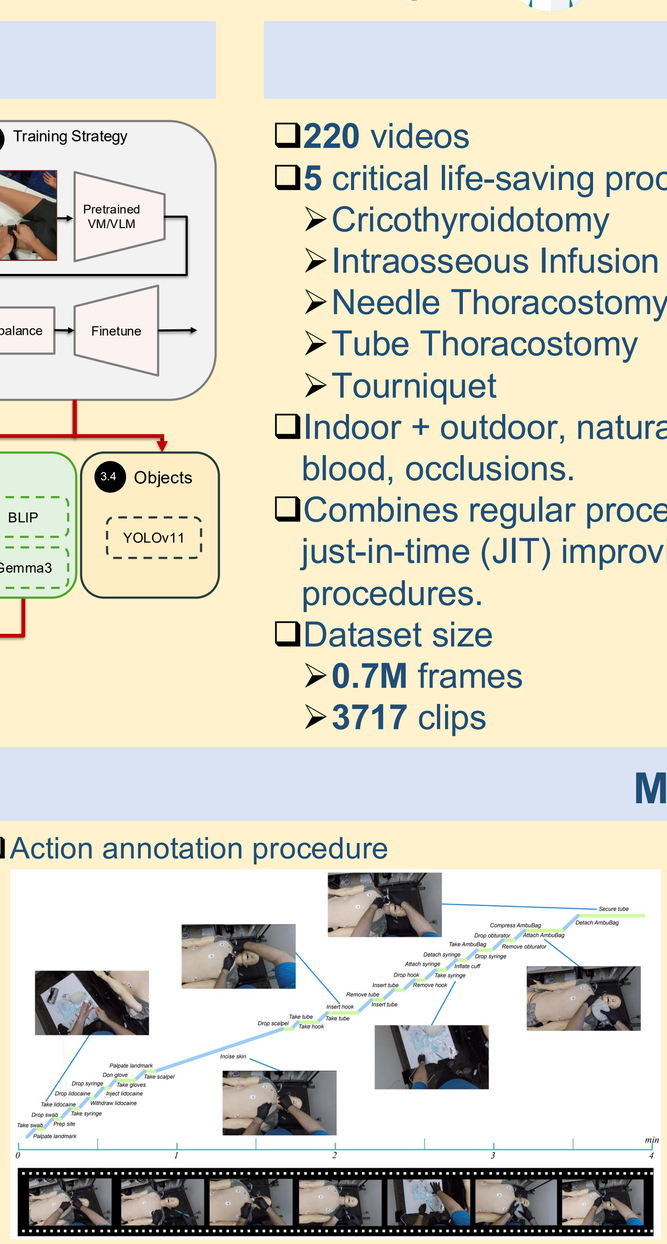
Qwen2.5-VL

Just-in-time Procedures

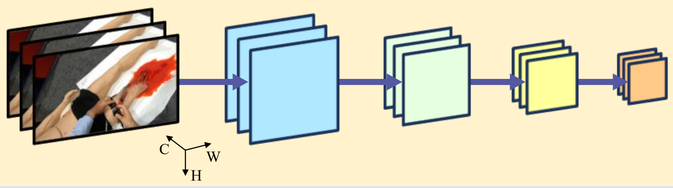
- A key component of the dataset focuses on JIT procedures, where responders must use improvised and non-traditional tools due to resource limitations.
- Examples: Using a belt and screwdriver for a tourniquet, or scissors to expand an incision for a tube thoracostomy.
- Challenge: These scenarios are difficult for AI models trained only on standard procedures.
- Finding: Training on a combined dataset (regular + JIT) significantly improves model performance and generalization, highlighting the need for diverse training data.

Methods

- Action annotation procedure



- Best Performing Model for Action Recognition and Action Anticipation – MViTv2



- Action annotation accuracy

$$t_a = \frac{1}{n_r + 1}\left(t_o + \sum_{i=1}^{n_r} t_{r_i}\right)$$

$$t_{\text{start}} = \max(t_{os}, t_{as}), \quad t_{\text{end}} = \min(t_{oe}, t_{ae})$$

$$p_i = \frac{t_{\text{end}} - t_{\text{start}}}{t_{ae} - t_{as}}$$

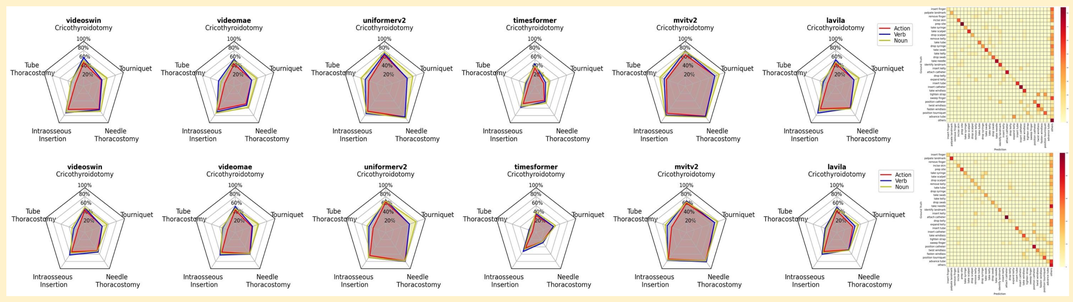
$$\text{Accuracy} = \frac{\sum_{i=1}^{n} p_i (t_{ae} - t_{as})}{\sum_{i=1}^{n} (t_{ae} - t_{as})}$$

t_a	Actual timestamp (average of annotator and reviewers)	t_o	Timestamp from the primary annotator
n_r	Number of reviewers	t_{r_i}	Timestamp from reviewer (i)
t_{as}, t_{ae}	Actual start and end times	t_{os}, t_{oe}	Original start and end times from annotator
		p_i	Annotation accuracy for clip (i)

- Hands & Objects
  - Bounding boxes
  - Format: COCO, YOLO

Benchmark Results

- Action Recognition
  - Best model (MViT v2) achieved 65% Top-1 accuracy on regular procedures.
  - Significant performance drop on JIT when trained only on regular (down to 14%).
  - Combined training (regular + JIT) restored performance to ~64% Top-1, proving the value of dataset diversity.
- Action Anticipation
  - Harder than recognition; Top-1 accuracy <10% on JIT when not trained jointly.



Train	Test	VideoSwin		TimeSFormer		VideoMAE		Uniformer v2		MViT v2		LaViLa	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Action Recognition													
Regular	Regular	45.10	74.52	31.91	62.81	43.34	71.89	60.47	85.65	65.59	89.75	42.52	68.81
	JIT	3.85	15.38	0.51	5.77	5.77	17.31	8.65	21.15	14.49	35.20	0.96	9.62
	Combined	34.60	66.71	27.70	55.27	38.37	64.68	53.62	77.13	58.58	82.08	38.25	60.99
Combined	Regular	44.51	73.35	29.42	63.69	48.61	73.06	60.32	84.19	66.47	89.17	40.17	66.67
	JIT	39.42	70.19	32.69	58.65	44.23	65.38	53.85	80.77	50.96	90.38	37.71	63.84
	Combined	43.84	72.94	29.86	63.02	48.03	72.05	59.47	83.74	64.42	88.82	39.53	66.02

Train	Test	VideoSwin		TimeSFormer		VideoMAE		Uniformer v2		MViT v2		LaViLa	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Action Anticipation													
Regular	Regular	39.88	69.24	28.44	59.97	41.42	69.24	56.25	84.70	60.12	87.02	38.79	67.85
	JIT	3.16	7.37	3.16	7.37	2.11	8.42	5.26	23.16	8.42	22.11	1.05	8.42
	Combined	35.18	61.32	25.20	53.23	36.39	61.46	49.73	76.82	53.50	78.71	33.96	60.24
Combined	Regular	41.89	69.09	26.74	61.21	44.05	69.40	57.95	83.77	60.74	86.56	40.03	67.58
	JIT	36.84	63.16	24.21	56.84	33.68	68.42	47.37	81.05	48.42	86.32	35.64	61.96
	Combined	41.24	68.33	26.42	60.65	42.72	69.27	56.60	83.42	59.16	86.52	39.65	66.73

- VQA
  - BLIP and Florence-2 reached ~88% accuracy.
  - Larger models (Qwen, LLaVA) were competitive but did not consistently outperform smaller, specialized models.

Model	Accuracy (%)	Parameters	Vision	Language
ViLT-B/32	79.88	87M	Transformer-based	BERT
BLIP-base	88.64	224M	ViT	BERT
Florence-2-base	87.86	230M	DaViT	Transformer-based
Qwen2.5-VL	83.29	7B	CLIP-based ViT	Qwen2.5 LLM
LLaVa-v1.6	85.47	7B	CLIP-based ViT	Vicuna
Gemma3	72.04	4B	SigLIP	Gemma 3 LLM

📌 **Takeaway: Trauma THOMPSON establishes clear baselines, reveals the limitations of current models in austere settings, and points to the need for multi-task and generalizable AI.**

Acknowledgement & Disclaimer

This work was supported by the US Army Medical Research and Development Command under Contract No. W81XWH21C0119 and by the National Science Foundation under Grant NSF #2140612.
The views expressed are those of the author(s) and do not reflect the official policy of the Department of the Army, the Department of Defense or the U.S. Government. The investigators have adhered to the policies for protection of human subjects as prescribed in 45 CFR 46.