The Trauma THOMPSON Dataset for Real-World Emergency AI

Yupeng Zhuo[1]
zhuoy@purdue.edu
Eddie Zhang[1]
eddie.r.zhang@gmail.com
Xiangchen Yu[1]
yu1335@purdue.edu
Aditya Pachpande[1]
apachpan@purdue.edu
Andrew Wallace Kirkpatrick[2]
andrew.kirkpatrick@albertahealthservices.ca
Kyle Couperus[3]
kcoup@uw.edu
Jessica Mckee[2]
jessicamckee05@gmail.com
Juan Wachs[1]
jpwachs@purdue.edu

[1] Edwardson School of Industrial Engineering
Purdue University
West Lafayette, Indiana, USA

[2] Departments of Critical Care and Surgery
University of Calgary
Foothills Medical Centre
Calgary, Alberta, Canada

[3] The Geneva Foundation
Tacoma, Washington, USA

Abstract

We introduce the Trauma THOMPSON dataset, a large-scale resource aimed at advancing artificial intelligence for real-time decision support in emergency and austere medical environments. The dataset consists of 220 unscripted, egocentric recordings of five emergency procedures, including a diverse set of "just-in-time" (JIT) life-saving interventions carried out under resource-limited conditions. These scenarios capture the realities of humanitarian and field medicine, where practitioners must frequently adapt or improvise beyond standard protocols. To enable richer visual understanding, we provide two new layers of fine-grained annotations: bounding boxes for critical medical instruments and supplies, and detailed hand annotations to support tracking and surgical skill assessment. Together, these enable research on spatiotemporal reasoning, interaction modeling, and the development of AI copilots capable of interpreting and assisting with complex procedures in real time. The dataset supports multiple benchmark tasks, including action recognition, action anticipation, visual question answering (VQA), and object detection. We benchmark state-of-the-art models across these tasks, highlighting both current strengths and the challenges of building reliable AI systems for deployment in the field. The dataset is publicly available at https://doi.org/10.7910/DVN/V5BTRU and provides a foundation for the creation of intelligent systems that can assist frontline medical personnel.

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

1 Introduction

Delivering effective medical care in remote, disaster, and combat settings is particularly challenging due to shortages of trained personnel, limited resources, and unreliable communication networks [1]. In such environments, first responders often possess only basic training yet must manage complex cases under severe constraints, which increases the likelihood of poor patient outcomes [2, 3]. Standard medical equipment is also frequently unavailable, forcing caregivers to improvise with everyday items. For instance, in the absence of proper bandages, torn clothing may be the only means of controlling severe bleeding. Prior research has demonstrated that "just-in-time" (JIT) training can enhance both skill acquisition and patient outcomes [4, 5]. Despite this, there remains a lack of data to guide the development of Artificial Intelligence (AI) assistants tailored for these critical contexts.

Several studies have suggested the use of AI assistants as copilots to support trainees and mentees in medical education and practice [6, 7, 8, 9, 10, 11, 12, 13, 14]. To advance this vision in low-resource environments and overcome the current lack of suitable data, we present the Trauma THOMPSON dataset (TTD). This dataset consists of egocentric video clips with detailed action annotations and accompanying benchmarks, designed to drive research on AI copilots for emergency and resource-limited scenarios. To our knowledge, TTD is the first dataset of its kind in terms of scale, diversity of settings, and the complexity of challenges it addresses. An overview of the experimental pipeline is shown in Figure 1. In brief, this work contributes a new dataset, annotations, and benchmark evaluations to support the development of next-generation AI medical assistants.

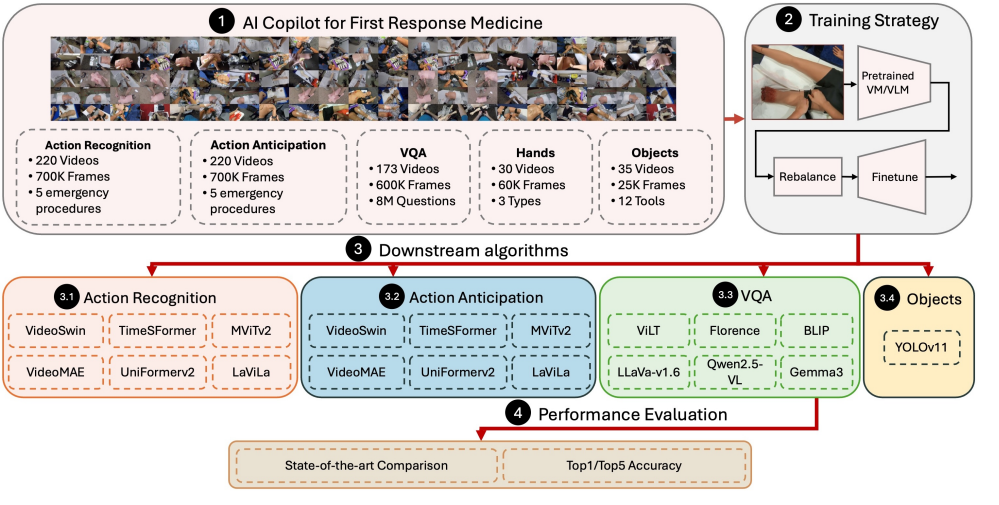

Figure 1: Overview of the experimental pipeline.

- This work introduces a pioneering egocentric view dataset for operational medicine, designed to assist field medics in properly performing emergency procedures within resource-limited environments. It contains annotated video clips covering five different unscripted life-saving skills.
- The dataset uniquely features procedures performed with both standard medical equipment and improvised "just-in-time" (JIT) procedures that use everyday objects. This

presents a significant challenge for AI, testing a model's ability to generalize and apply complex reasoning for medical skill transfer between standard and improvised scenarios.

- We establish benchmarks for action recognition and anticipation, which are crucial for predicting the therapeutic actions required in humanitarian and resuscitative medicine. This work is intended to serve as a foundational piece for creating AI copilots to support medics and first responders.

- We have created fine-grained annotations for visual question answering (VQA), object detection, and hand tracking. We also provide benchmarks for object detection and for VQA to showcase its potential as a clinical decision support (CDS) tool that can guide caregivers through a natural dialogue.

2 Related Work

2.1 Egocentric activity recognition and surgical datasets

Video understanding has seen dramatic advances due to the introduction of action classification benchmarks such as UCF101 [15], HMDB51 [16], Kinetics [17], Something-Something [18], and AVA [19], which mostly consist of short videos focusing on a single action per clip and aim to recognize daily activities. Nonetheless, these datasets may lack the spontaneity, progression, and multi-tasking that occur in real-life situations due to their scripted nature. As a result, research has shifted focus to first-person vision, which delivers activities from a unique viewpoint. For instance, [20] developed a dataset that includes 20 participants and encompasses 10 hours of activities of daily living (ADL) videos. [21] created EGTEA Gaze+ with wearable cameras, which is an egocentric ADL dataset of 28 hours of cooking activities from 86 distinct sessions involving 32 subjects. [22] curated the EPIC-KITCHENS dataset, which is a large-scale egocentric video dataset with 100 hours of cooking actions recorded by 32 participants.

There are multiple robotic surgery datasets for the computation of proficiency, skill level, knowledge acquisition, and performance [23, 24, 25]. Moreover, instructional videos for life-saving skills have been proposed for the purpose of training AI algorithms [26]. However, such datasets were collected in controlled settings using both simulation and planned surgical procedures.

Figure 2: Examples of JIT procedures under resource limited settings. The first row shows performing cricothyroidotomy with scissors. The second row shows performing tourniquet with a screwdriver and a belt.

Table 1: Comparison of Trauma THOMPSON to the related egocentric and medical datase

Dataset	Ego	Med	Austere	Frames	No. Act	Participants	No. Envs
Trauma THOMPSON, 2025	X	X	X	0.7M	162	12	15
EPIC-KITCHENS, 2018 [22]	X			11.5M	149	32	32
Ego-SLD, 2025 [27]	X			892 videos	16	12	12
BEOID, 2014 [28]	X			0.1M	34	5	1
GTEA, 2011 [29]	X			0.4M	42	13	1
CMU-MMAC, 2008 [30]	X			0.2M	31	16	1
ADL, 2012 [20]	X			1.0M	32	20	20
ESAD, 2020 [31]		X		0.03M	21	4	4
CholecT50, 2022 [32]		X		0.1M	100	13	13
MedVidCL, 2023 [26]		X		1489 Videos	0	>100	>100
MRAO, 2021 [33]		X		480 Videos	10	16	2
MISAW, 2021 [34]		X		27 Videos	17	6	1
PSI-AVA, 2022 [35]		X		8 Videos	167	3	1
PETRAW, 2023 [36]		X		150 Videos	6	4	2

2.2 Our work: Egocentric operational medicine dataset

Table 1 compares the TTD to common egocentric view and medical instructional datase and presents key metrics that distinguish the TTD as the rst egocentric view medical in structional dataset with per-frame annotations.

The TTD has a similar structure as other egocentric datasets for action recognition ar anticipation, such as EPIC KITCHENS [22], GTEA [29] and EGTEA Gaze+ [21], and Charades-Ego [37], with the following caveats. Firstly, the hands are not always visible o distinguishable due to arti cial blood, occlusions, and multiple limbs, which makes it more challenging for detection and tracking. Secondly, some of the videos are taken outdoo and "in the wild", increasing the complexity due to uncontrolled lighting. Thirdly, mistakes require rewinding and re-doing, or stopping short while completing the procedures, leadin to a high variability in style and performance time. Lastly, as opposed to existing dataset for surgical guidance and instruction, which rely on a xed set of tools common to general surgery, our dataset is subject to emergency settings. The performers are rst responde medics, and surgeons, and the procedures were often conducted using improvised tools (e scissors for cricothyroidotomy, belt for a tourniquet) to replicate a resource-limited setting as shown in Figure 2. This poses a challenge for algorithms that rely on object detecto as priors for activity recognition, as the objects are not known ahead of time in emerger settings.

3 Dataset

3.1 Procedure identi cation

The development of the TTD involved a team of experts with experience in deployed settings such as surgeons, critical care physicians, and emergency medicine physicians, who crea a list of essential procedures for prolonged casualty care (PCC), such as cricothyrotomy a tourniquet application. Additionally, a focus group of 15-30 subject matter experts (SMEs) determined a consensus on the content and best practices for the TTD. This information w used to identify a nal list of procedures for the TTD, which includes cricothyroidotomy, in- traosseous infusion, tourniquet, needle thoracostomy, and tube thoracostomy. The collecti of procedures and settings are described in the following section.

3.2 Data collection

We focused on capturing natural, unscripted life-saving intervention (LSI) procedures fr the rst-person perspective, which involves operating a medical tool, searching for an ite changing one's mind, and encountering unexpected problems. The videos were recorde 1080p using head-mounted cameras (GoPro, Hero7, San Mateo, California) to capture person views lmed across various simulation models and environments. Surgeons w the cameras on their heads and adjusted the angle to 20-30° relative to the forehea optimal video collection. The hands were centered in frame during procedures for be visualization.

The dataset was also enhanced through the inclusion of videos capturing JIT proced involving improvised, non-traditional equipment. Videos were obtained of users performi improvised tourniquets (utilizing belts or clothing and a screwdriver), tube thoracosto (utilizing scissors for incision and expansion of thoracostomy and a screwdriver to gui insertion of the tube), needle cricothyroidotomy (replacing standard incision/tube with needle for emergent airway management), and manual intraosseous needle placement (needle driver is not available or functional).

3.3 Annotation pipeline

The annotations in the dataset consist of start timestamps, end timestamps, and action pressed as verb-noun pairs for corresponding video clips. The expected output for testi the labels for the action, verb, and noun. Medical professionals were responsible for ar tating the data and providing the timestamps and actions for each procedural step. To re the possibility of errors in time stamping and video segmentation, the annotations underw peer review.

3.4 Data quality assurance

To ensure annotation accuracies, the actions in each procedure are annotated by three m professionals. One person annotates and the other two people review the generated ar tions. As the annotations are estimates and no precise way to ensure an absolute times for each procedure, we propose a method to compute the annotation accuracy. Let the actual timestamp and de ned as the average of timestamps from the annotator and reviewers. n_r is the number of reviewers. t_{ri} is the timestamp from reviewer i. t_o is the timestamp from the annotator. $t_a = \frac{1}{n_r+1}(\sum_{i=1}^{n_r} t_{ri} + t_o)$. t_{as} and t_{ae} denote the actual start and end of each clip. t_{os} and t_{oe} denote the original start and end by the annotator. The annotati accuracy of each clip is computed as the overlapping time between the original and ac timestamps divided by the actual clip duration. To compute the overlapping time, we de $t_{start} = \max(t_{os}, t_{as})$ and $t_{end} = \min(t_{oe}, t_{ae})$. The clip accuracy p_i is computed as $\frac{t_{end} - t_{start}}{t_{ae} - t_{as}}$.

The average annotation accuracy is computed as $acc = \frac{\sum_{i=1}^{n}(p_i \cdot (t_{ae} - t_{as}))}{\sum_{i=1}^{n}(t_{ae} - t_{as})}$.

3.5 VQA annotations

The medical VQA is derived from the egocentric video dataset and includes additional notations that contain questions and corresponding plausible answers. Each question i secondary annotations contains 3 to 5 potential answers. For example, Q: What limb is

injured? A: Right arm; Q: Where is the catheter insderted? A: There is no catheter; A: Is there any bleeding? A: No.

3.6 Hand and object annotations

To annotate high-quality bounding boxes ef ciently, the human-in-the-loop approach is adop which combines both manual annotation and automatic tracking. The bounding boxes a created by manual selections of hands and objects in the videos every 10-30 frames and au matically annotated by CSRT trackers [38] between selections. Left hand, right hand, and 1 medical tools are annotated. Teaching vision language models (VLMs) to track hands ar recognize objects is clinically signi cant, especially in high-stakes medical environments. Accurate hand tracking enables AI assistants to assess procedural skills in real-time, offe ing immediate feedback on bimanual coordination and task execution [39, 40]. Meanwhile integrating object detection with natural language understanding allows clinicians to ask A assistants where speci c tools are located, reducing cognitive load and minimizing the ris of human error. Various VLMs have demonstrated object detection capability [41], such a Florence-2 [42] and F-VLM [43], highlighting the potentials to train uni ed VLMs that can perform various vision tasks to assist medical procedures.

3.7 Dataset classes distribution

The dataset comprises 220 videos demonstrating 5 medical procedures and contains 37 fully annotated video clips. For action classes, the distribution is in accordance to real-worl scenarios, leading to a long-tailed dataset. The regular procedure includes 42 verb class 42 noun classes, and 124 action classes, while the JIT procedure includes 28 verb classes noun classes, and 86 action classes.

3.8 Annotation accuracy statistics

Due to the large volume of the dataset, the reviewers were requested to randomly revie 60 videos in the dataset according to the above-stated instructions. The temporal accura is 99.4%, the label accuracies of actions, verbs, and nouns are 97.2%, 97.2%, and 97.7 respectively.

4 Benchmark Results

4.1 Action recognition and action anticipation

Evaluation setups and metrics The dataset is split into the train and test sets by 80% and 20% of the data, respectively, for both the regular and JIT procedures. We trained a algorithms on the regular or combined (regular+JIT) set and tested them in three categorie regular alone, JIT alone, and combined setting. Each model was trained on a GeForce RTX 4090 Ti. A class-agnostic approach was used to assess accuracy [44] of the models. T Top1 and Top5 accuracies for verb, noun, and action (verb + noun) were evaluated.

Action recognition We trained on vision models (VMs) and VLMs for action recogni- tion using the TTD on both the regular and JIT procedures. All models were pretrain and then netuned on the TTD. TTD is a long-tailed dataset, so random oversampling w adopted to rebalance the dataset. Table 2 presents a performance comparison of va action recognition models evaluated on different train-test con gurations. Six models a assessed: VideoSwin [45], TimeSFormer [46], Uniformer v2 [47], VideoMAE [48], MViT v2 [49], LaViLa [50]. Each model's performance is measured using Top1 and Top5 accura metrics under three different train-test scenarios: regular, JIT, and combined.

Table 2: Accuracy (%) comparison of action recognition models on different train-test s tings. Purple for lower values, while green for higher values.

Train	Test	VideoSwin		TimeSFormer		VideoMAE		Uniformer v2		MViT v2		LaViLa	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
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

In the rst scenario, when models are trained and tested on regular data, MViT achieves the highest accuracy, with 65.59% Top1 and 89.75% Top5. Uniformer v2 follo closely, with Top1 and Top5 accuracies of 60.47% and 85.65%, respectively. In contra TimeSFormer has the lowest performance, achieving only 31.91% Top1 and 62.81% T accuracy. When tested on JIT data with regular training, all models experience a substa drop in performance. MViT v2 remains the best performer with a Top1 accuracy of 14.49 but other models, like TimeSFormer and LaViLa, perform poorly, with Top1 accuracies b low 1%. In the combined test setting with regular training, a similar trend is observed, whe MViT v2 continues to outperform other models, achieving 58.58% Top1 and 82.08% To accuracy, and TimeSFormer remains the least effective model. This indicates that reg training does not generalize well to JIT or combined testing scenarios for most models.

In the second training scenario, where models are trained on combined data, their formance improves signi cantly in the JIT test setting. MViT v2 achieves the highest Top accuracy of 90.38%, while Uniformer v2 performs best in terms of Top1 accuracy of 53.85 Compared to the regular training, this training strategy signi cantly improves performan on the JIT test data for all models, highlighting the bene ts of incorporating diverse trainin data. For the combined test setting under combined training, MViT v2 continues to perfo the best with a Top1 accuracy of 64.42% and a Top5 accuracy of 88.82%. Uniformer v2 a VideoMAE also achieve high accuracies in this setup, while TimeSFormer remains the le effective across all metrics and scenarios.

Figure 3a presents the confusion matrix of the best performing model MViT v2 on th action recognition task. The dark color in the diagonal direction indicates high predicti accuracy of the model. The labels are arranged by action class frequency, with more freq classes on the top and less frequent at the bottom. It can be seen that there are some dark in the lower part of the gure, indicating the dif culty of the model to predict less frequen classes. Figure 4 illustrates the detailed top 1 accuracy of verb, noun, action of recogniz the ve types of emergency procedures for each model. The similarity of shapes in the ra charts indicates the coherence in the performance of the models for each procedure. It be seen that the best performing algorithm does not triumph in all procedures.