MELBA Journal Machine Learning for Biomedical Imaging

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

Abstract

We present the Trauma THOMPSON dataset and benchmarks designed to advance artificial intelligence research for real-time decision support in emergency and austere medical environments. The dataset contains 220 unscripted egocentric videos of five emergency procedures, including a diverse collection of "just-in-time" (JIT) life-saving interventions performed under resource-constrained conditions. These JIT scenarios more closely reflect the realities of humanitarian and field-based operational medicine, where standard protocols must often be adapted or creatively executed. To support deeper visual understanding, we introduce two new layers of fine-grained annotations: object detection labels for critical medical instruments and supplies and hand annotations to facilitate hand tracking and surgical skill assessment. These additions enable new research directions in spatiotemporal reasoning, interaction modeling, and AI copilots that interpret and guide complex procedures in real time. The Trauma THOMPSON dataset includes benchmark tasks in action recognition, action anticipation, visual question answering (VQA), object detection, and hand localization. We evaluate state-of-the-art models across these tasks, identifying current strengths and open challenges in developing robust AI for field-deployable decision-making. The dataset is available at `https://github.com/zhuoyp/TTD`, and it can serve as a foundation for building intelligent systems that assist frontline caregivers.

Keywords

Emergency Procedures, Improvisation, Action Recognition, Action Anticipation, Hand Tracking, Object Detection

Article informations

`https://doi.org/https://doi.org/10.59275/j.melba.2025-5ce1`

Volume 3, Published 12/2025

Special issue: MICCAI Open Data special issue

Guest editors: Martijn Starmans, Apostolia Tsirikoglou, Lidia Garrucho Moras, Kaouther Mouheb

1. Introduction

Providing high-quality medical care in remote, disaster-stricken, and combat environments presents significant challenges due to limited medical expertise, scarce resources, and unreliable connectivity (Wachs, Kirkpatrick, and Tisherman 2021). In these conditions, first responders often have minimal training yet must handle complex medical cases with constrained resources, heightening the risk of poor patient outcomes (Stewart and Bird 2022; Shackelford et al. 2021). Furthermore, standard medical tools and resources are frequently unavailable under remote environments, and using daily objects to perform emergency procedures becomes the inevitable choice. For example, if someone is bleeding heavily, tearing clothing to help stop the bleeding will be the only option when medical bandages are not easily available. Previous studies have shown that "just-in-time" (JIT) training can improve learning and patient outcomes (Patocka et al. 2024; Branzetti et al. 2017). However, few data on training Artificial Intelligence (AI) assistants are available in this domain.

AI assistants have been proposed to act as copilots for a mentee or trainee by multiple studies (Bahl 2020; X. Liu et al. 2018; Al-Antari 2023; Dilsizian and Siegel 2014; Hamet and Tremblay 2017; Dinh et al. 2023; Mirchi et al. 2020; Vannaprathip et al. 2025; Caballero et al. 2025). To train AI medical assistants for low-resource settings and address the gap in data availability, we introduce the Trauma THOMPSON dataset (TTD), which is a collection of video clips with action annotations, and benchmarks to encourage research and development of AI copilots for

resource-constrained and emergency settings.

TTD was used as the basis for the "MICCAI 2023 Trauma THOMPSON Challenge" (Zhuo, Kirkpatrick, et al. 2025; Zhuo, W. Kirkpatrick, et al. 2025), which focused on action recognition and action anticipation of regular emergency procedures. This paper presents a comprehensive release of the TTD and presents the following innovations beyond our previous work. Specifically, it expands the dataset by including JIT procedures and adding new annotations for hands, objects, and visual question answering (VQA), along with new benchmark results. To the best of our knowledge, TTD is the first of its kind in terms of scale, settings, challenges, and applicability. Figure 1 shows an overview of the experimental pipeline of this study. In summary, this paper makes the following contributions.

- We created the first egocentric view dataset of operational medicine that assist field medics to properly perform emergency care procedures in resource-constrained settings. This dataset includes annotated video clips corresponding to 5 unscripted procedures for life-saving skills.
- The dataset contains both regular procedures performed with standard medical tools and JIT procedures performed with daily objects. By training on regular procedures and testing on JIT procedures, we introduce a very challenging scenario that tests an AI model's complex reasoning and understanding of medical skill transfer and domain generalization.
- We provide benchmarks for action recognition and anticipation to predict the therapeutic actions required for humanitarian medicine and resuscitative care. This dataset is intended to act as an essential piece for developing copilots for medics and first responders.
- We created secondary annotations for VQA, hand tracking, and object detection. We provide benchmarks for the VQA task to illustrate how it can be potentially used as a clinical decision support (CDS) tool to assist caregivers through natural dialogue throughout the diagnostic process.

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 (Soomro, Zamir, and Shah 2012), HMDB51 (Kuehne et al. 2011), Kinetics (Kay et al. 2017), Something-Something (Goyal et al. 2017), and AVA (Gu et al. 2017), 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, (Pirsiavash and Ramanan 2012) developed a dataset that includes 20 participants and encompasses 10 hours of activities of daily living (ADL) videos. (Yin Li, M. Liu, and Rehg 2020) 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. (Damen et al. 2018) 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 (Gao et al. 2014; Tao et al. 2012; Gonzalez et al. 2021). Moreover, instructional videos for life-saving skills have been proposed for the purpose of training AI algorithms (Gupta, Attal, and Demner-Fushman 2023). However, such datasets were collected in controlled settings using both simulation and planned surgical procedures.

2.2 Our work: Egocentric operational medicine dataset

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

The TTD has a similar structure as other egocentric datasets for action recognition and anticipation, such as EPIC KITCHENS (Damen et al. 2018), GTEA (Fathi, Ren, and Rehg 2011) and EGTEA Gaze+ (Yin Li, M. Liu, and Rehg 2020), and Charades-Ego (Sigurdsson et al. 2018), with the following caveats. Firstly, the hands are not always visible or distinguishable due to artificial blood, occlusions, and multiple limbs, which makes it more challenging for detection and tracking. Secondly, some of the videos are taken outdoors 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, leading to a high variability in style and performance time. Lastly, as opposed to existing datasets for surgical guidance and instruction, which rely on a fixed set of tools common to general surgery, our dataset is subject to emergency settings, as shown in Figure 2. The performers are first responders, medics, and surgeons, and the procedures were often conducted using improvised tools (e.g. shirt for a tourniquet, a pocket knife for a cricothyroidotomy) to replicate a resource-limited setting. This poses a challenge for algorithms that rely on object detectors as priors for activity recognition, as the objects are not known ahead of time in emergent settings.

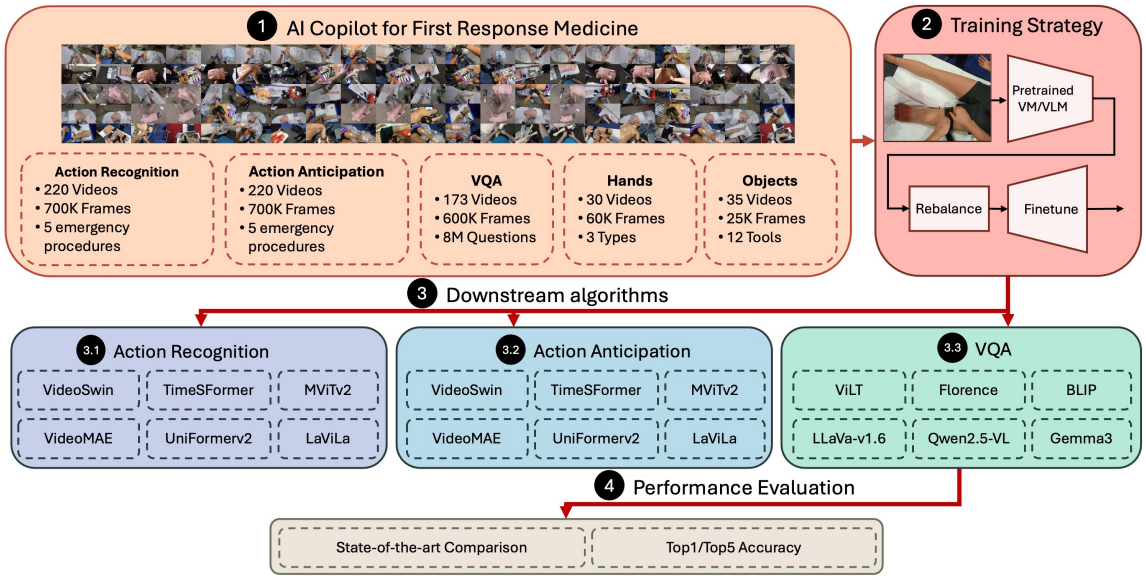


Figure 1: Overview of the experimental pipeline.

Table 1: Comparison of Trauma THOMPSON to the related egocentric and medical datasets. "Ego" denotes egocentric dataset, and "Med" denotes medical dataset.

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

3. Dataset

3.1 Procedure identification

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 created a list of essential procedures for prolonged casualty care (PCC), such as cricothyrotomy and 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 was used to identify a final list of procedures for the TTD, which includes cricothyroidotomy, intraosseous infusion, tourniquet, needle thoracostomy, and tube thoracostomy. The collection 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 from the first-person perspective, which involves operating a medical tool, searching for an item, changing one's mind, and encountering unexpected problems. The videos were recorded at 1080p using head-mounted cameras (GoPro, Hero7, San Mateo, California) to capture first-person views filmed across various simulation models and environments. Surgeons wore the cameras on their heads and adjusted the angle to 20-30° relative to the forehead for optimal video collection. The hands were centered in frame during procedures for better visualization.

The dataset was also enhanced through the inclusion of videos capturing JIT procedures involving improvised, non-traditional equipment. Videos were obtained of users performing improvised tourniquets (utilizing belts or clothing and a screwdriver), tube thoracostomy (utilizing scissors for incision and expansion of thoracostomy and a screwdriver to guide insertion of the tube), needle cricothyroido-

Figure 2: Examples of procedure video clips.

tomy (replacing standard incision/tube with a needle for emergent airway management), and manual intraosseous needle placement (when needle driver is not available or functional).

3.3 Annotation pipeline

The annotations in the dataset consist of start timestamps, end timestamps, and actions expressed as verb-noun pairs for corresponding video clips. The expected output for testing is the labels for the action, verb, and noun. Medical professionals were responsible for annotating the data and providing the timestamps and actions for each procedural step. To reduce the possibility of errors in time stamping and video segmentation, the annotations underwent peer review.

3.4 Data quality assurance

To ensure annotation accuracies, the actions in each procedure are annotated by three medical professionals. One person annotates and the other two people review the generated annotations. As the annotations are estimates and no precise way to ensure an absolute timestamp for each procedure, we propose a method to compute the annotation accuracy. Let t_a be the actual timestamp and defined as the average of timestamps from the annotator and the 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 annotation accuracy of each clip is computed as the overlapping time between the original and actual timestamps divided by the actual clip duration. To compute the overlapping time, we define $t_{start} = \max(t_{os}, t_{as})$ 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 $\text{acc} = \frac{\sum_{i=1}^{n}(p_i*(t_{ae}-t_{as}))}{\sum_{i=1}^{n}(t_{ae}-t_{as})}$.

3.5 VQA annotations

The medical VQA is derived from the egocentric video dataset and includes additional annotations that contain questions and corresponding plausible answers. Each question in the 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 efficiently, the human-in-the-loop approach is adopted, which combines both manual annotation and automatic tracking. The bounding boxes are created by manual selections of hands and objects in the videos every 10-30 frames and automatically annotated by CSRT trackers (Lukežič et al. 2018) between selections. Left hand, right hand, and 12 medical tools are annotated. Teaching vision language models (VLMs) to track hands and recognize objects is clinically

significant, especially in high-stakes medical environments. Accurate hand tracking enables AI assistants to assess procedural skills in real-time, offering immediate feedback on bimanual coordination and task execution (Azari et al. 2019; Mackenzie et al. 2021). Meanwhile, integrating object detection with natural language understanding allows clinicians to ask AI assistants where specific tools are located, reducing cognitive load and minimizing the risk of human error. Various VLMs have demonstrated object detection capability (Feng et al. 2025), such as Florence-2 (Xiao et al. 2023) and F-VLM (Kuo et al. 2022), highlighting the potentials to train unified 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 3717 fully annotated video clips. For action classes, the distribution is in accordance to real-world scenarios, leading to a long-tailed dataset. The regular procedure includes 42 verb classes, 42 noun classes, and 124 action classes, while the JIT procedure includes 28 verb classes, 32 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 review 60 videos in the dataset according to the above-stated instructions. The temporal accuracy 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 all algorithms on the regular or combined (regular+JIT) set and tested them in three categories: 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 of the models (H. Zhao et al. 2019). The 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 recognition using the TTD on both the regular and JIT procedures. All models were pretrained and then finetuned on the TTD. TTD is a long-tailed dataset, so random oversampling was adopted to rebalance the dataset. Table 2 presents a performance comparison of various action recognition models evaluated on different train-test configurations. Six models are assessed: VideoSwin (Z. Liu et al. 2022), TimeSFormer (Bertasius, Wang, and Torresani 2021), Uniformer v2 (K. Li et al. 2022), VideoMAE (Tong et al. 2022), MViT v2 (Yanghao Li et al. 2022), LaViLa (Y. Zhao et al. 2022). Each model's performance is measured using Top1 and Top5 accuracy metrics under three different train-test scenarios: regular, JIT, and combined.

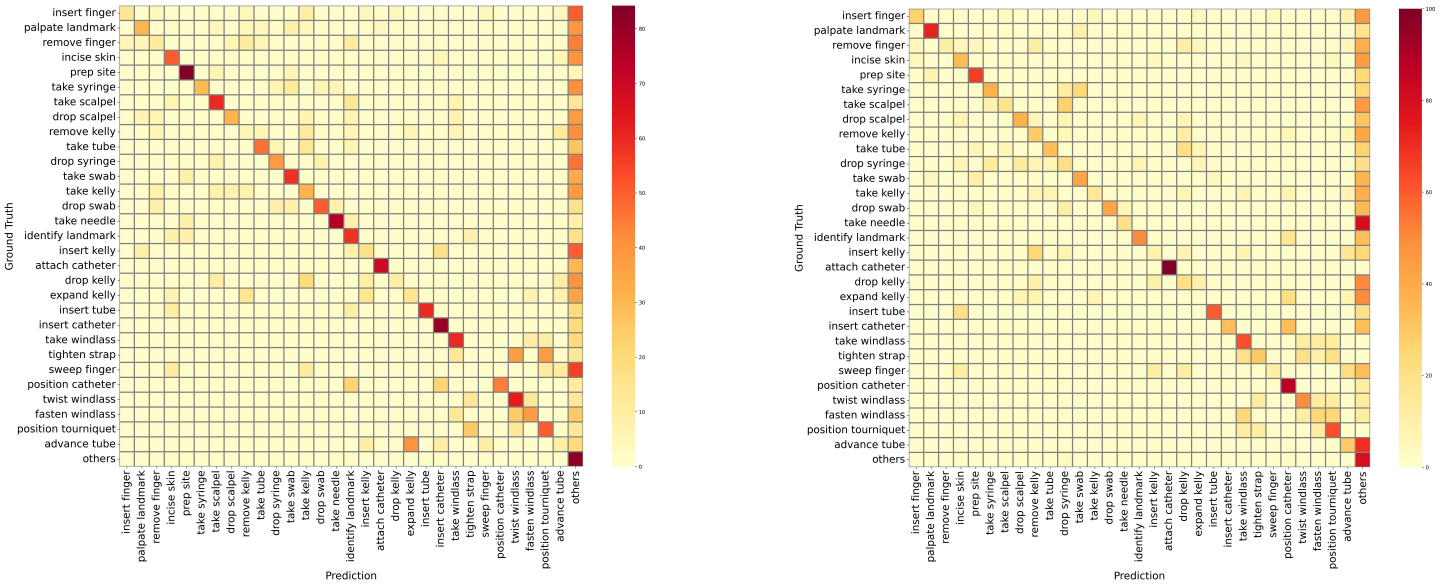
In the first scenario, when models are trained and tested on regular data, MViT v2 achieves the highest accuracy, with 65.59% Top1 and 89.75% Top5. Uniformer v2 follows closely, with Top1 and Top5 accuracies of 60.47% and 85.65%, respectively. In contrast, TimeSFormer has the lowest performance, achieving only 31.91% Top1 and 62.81% Top5 accuracy. When tested on JIT data with regular training, all models experience a substantial 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 below 1%. In the combined test setting with regular training, a similar trend is observed, where MViT v2 continues to outperform other models, achieving 58.58% Top1 and 82.08% Top5 accuracy, and TimeSFormer remains the least effective model. This indicates that regular 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 performance improves significantly in the JIT test setting. MViT v2 achieves the highest Top5 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 significantly improves performance on the JIT test data for all models, highlighting the benefits of incorporating diverse training data. For the combined test setting under combined training, MViT v2 continues to perform the best with a Top1 accuracy of 64.42% and a Top5 accuracy of 88.82%. Uniformer v2 and VideoMAE also achieve high accuracies in this setup, while TimeSFormer remains the least effective across all metrics and scenarios.

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

Table 2: Accuracy (%) comparison of action recognition models on different train-test settings. 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



(a) Action recognition.

(b) Action anticipation.

Figure 3: Confusion matrices of action recognition and action anticipation with MViT v2 on regular procedures.

Action Anticipation Table 3 presents the benchmarking results for action anticipation on the same six models. For the regular train-test setup, MViT v2 achieves the highest Top1 and Top5 accuracy of 60.12% and 87.02% , followed by Uniformer v2 with 56.25% Top1 and 84.70% Top5 accuracy. Other models, such as VideoSwin and VideoMAE, exhibit moderate performance, while TimeSFormer struggles with a Top1 accuracy of only 28.44%. When tested on JIT data with regular training, all models show big performance degradation, with MViT v2 achieving a maximum Top1 accuracy of only 8.42%. Uniformer v2 performs similarly, highlighting the challenges of generalizing anticipation models to JIT scenarios when trained on regular data. Under the combined test setting with regular training, MViT v2 again performs best, with a Top1 accuracy of 53.50% and a Top5 accuracy of 78.71%. Uniformer v2 closely follows, while TimeSFormer remains the weakest model. These trends are consistent with those observed in the regular test setting, underscoring the difficulty of training anticipation models for diverse scenarios using only regular procedure data.

The action anticipation results follow a similar trend to action recognition, with MViT v2 and Uniformer v2 outperforming other models across all scenarios. However, the absolute performance metrics for anticipation are slightly lower than those for recognition, particularly in the JIT test setting, where models experience a sharper decline in accuracy. This suggests that predicting future actions is inherently more challenging than recognizing current actions, especially in diverse or unexpected scenarios such as JIT procedures. The benefit of training on combined data is evident in both tasks, as it significantly enhances the models' generalization capabilities for regular and JIT settings.

Figure 3b shows the confusion matrix of the best performaning model MViT v2 on the action anticipation task. Similar performance has already been observed in action recognition, with less frequent classes being harder to classify for the model. Figure 5 illustrates the comparison of different models for action anticipation on the five emergency procedures. Similar plots are observed in action recognition, but with increased performance of the models

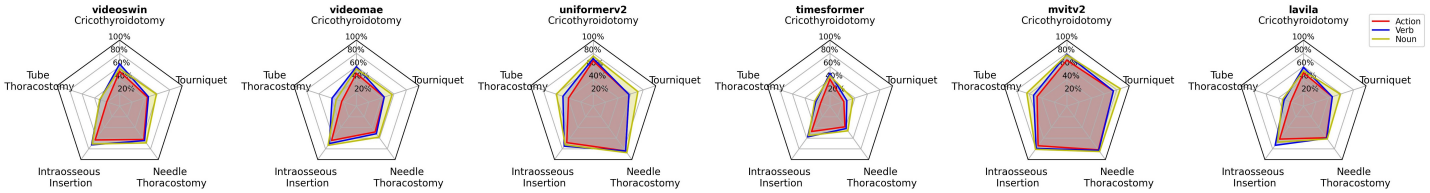


Figure 4: Action recognition Top 1 accuracies of verb, noun, and action by each type of procedure.

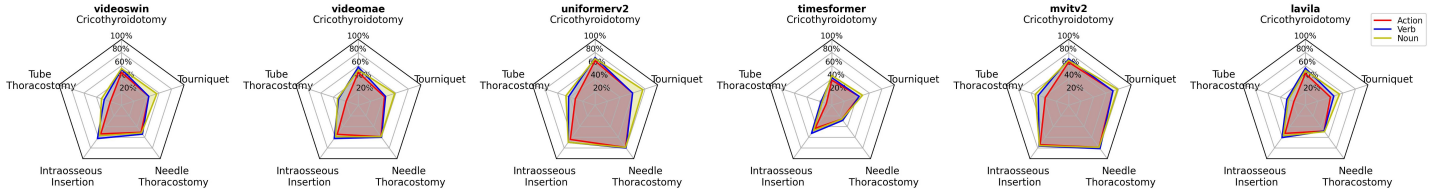


Figure 5: Action anticipation Top 1 accuracies of verb, noun, and action by each type of procedure.

Table 3: Accuracy (%) comparison of action anticipation models on different train-test settings. 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	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

compared to the action anticipation task.

4.2 VQA

Table 4 compares the performance of 6 finetuned VQA models of various sizes, ViLT-B/32 (W. Kim, Son, and I. Kim 2021), BLIP (J. Li et al. 2022), Florence2(Yuan et al. 2021), LLaVA-v1.6-7B (H. Liu, C. Li, Wu, et al. 2023; H. Liu, C. Li, Yuheng Li, et al. 2024), Qwen2.5-VL-7B (Bai et al. 2025), and Gemma3-4B (Gemma Team et al. 2025). LLaVA-v1.6, Qwen2.5-VL and Gemma3 are finetuned with the QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al. 2023) approach for efficient adaptation on a single GPU. BLIP achieved the highest accuracy of 88.64%, demonstrating strong VQA capabilities with a relatively moderate size. It was followed by Florence-2 with 87.86% accuracy. LLaVa-v1.6 achieved an accuracy of 85.57%, Qwen2.5-VL(7B) achieved an accuracy of 83.29%, and Gemma3 achieved an accuracy of 72.04%. ViLT-B/32 offers a lighter alternative with only 87 million parameters and an accuracy of 79.88%.

5. Conclusion

In this paper, we present the first egocentric dataset of five different LSI procedures for providing care in austere and adverse settings, along with benchmarks. We introduced various new challenges, such as hand occlusions, mistakes, and outdoors and field settings using improvised tools. Additionally, we created JIT procedures in the dataset and it is very challenging for VMs to make inferences in a zero-shot learning manner. Another characteristic of our dataset is that it is unscripted, leading to a significant variation between different performers, based on their experience in operational medicine. The TTD has simplified annotations (verb + noun). Future efforts will include allowing the actions to be annotated using complete sentences with detailed instructions and expanding the dataset with more varied procedural scenes, including additional improvised actions and broader context. Increasing dataset diversity could greatly enhance the performance of algorithms. Such efforts are critical to advancing the development of AI medical assistants in the complex domain of humanitarian medicine. Moreover, training unified VLMs that can solve all tasks in the TTD will greatly increase the clinical values of AI medical assistants.

Table 4: VQA model performance comparison.

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

Acknowledgments

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, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

Ethical Standards

The IRB approval was conducted by the Geneva team and obtained with the protocol number 223046.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

The Trauma THOMPSON dataset will be openly available for research and development purposes. Detailed instructions on how to process the dataset will be provided in the README file in our GitHub repository, including preprocessing steps and annotation formats. The source code for processing the videos, extracting relevant features, and building models will also be available. This will allow researchers to efficiently work with the dataset and implement their own machine learning pipelines. All users of the dataset will be required to fill out a consent form, after which the raw videos and annotations will be available for download. The dataset is hosted on Google Drive through a private link and access will be automatically granted upon submission of the consent form. To ensure consistent access and version control, we recommend that users begin from our official GitHub page, `https://github.com/zhuoyp/TTD`, which contains the most up-to-date instructions, links, and supporting code. The dataset is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.

References

Al-Antari, Mugahed A. (Feb. 12, 2023). "Artificial Intelligence for Medical Diagnostics—Existing and Future AI Technology!" In: *Diagnostics* 13.4, p. 688. ISSN: 2075-4418.

Azari, David P. et al. (Mar. 2019). "Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating". In: *Annals of Surgery* 269.3, pp. 574–581. ISSN: 0003-4932, 1528-1140.

Bahl, Manisha (Aug. 10, 2020). "Artificial Intelligence: A Primer for Breast Imaging Radiologists". In: *Journal of Breast Imaging* 2.4, pp. 304–314. ISSN: 2631-6110, 2631-6129.

Bai, Shuai et al. (Feb. 19, 2025). *Qwen2.5-VL Technical Report*. arXiv: 2502.13923[cs].

Bawa, Vivek Singh et al. (June 12, 2020). *ESAD: Endoscopic Surgeon Action Detection Dataset*. arXiv: 2006.07164[cs].

Bertasius, Gedas, Heng Wang, and Lorenzo Torresani (2021). *Is Space-Time Attention All You Need for Video Understanding?* Version Number: 4.

Branzetti, Jeremy B et al. (Nov. 2017). "Randomised controlled trial to assess the effect of a Just-in-Time training on procedural performance: a proof-of-concept study to address procedural skill decay". In: *BMJ Quality & Safety* 26.11, pp. 881–891. ISSN: 2044-5415, 2044-5423.

Caballero, Daniel et al. (Jan. 30, 2025). "Applications of Artificial Intelligence in Minimally Invasive Surgery Training: A Scoping Review". In: *Surgeries* 6.1, p. 7. ISSN: 2673-4095.

Damen, Dima (2014). *Bristol Egocentric Object Interactions Dataset*. In collab. with Walterio Mayol-Cuevas and Teesid Leelasawassuk.

Damen, Dima et al. (July 31, 2018). *Scaling Egocentric Vision: The EPIC-KITCHENS Dataset*. arXiv: 1804.02748[cs].

Dettmers, Tim et al. (May 23, 2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314[cs].

Dilsizian, Steven E. and Eliot L. Siegel (Jan. 2014). "Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment". In: *Current Cardiology Reports* 16.1, p. 441. ISSN: 1523-3782, 1534-3170.

Dinh, Alana et al. (Apr. 18, 2023). "Augmented Reality in Real-time Telemedicine and Telementoring: Scoping Review". In: *JMIR mHealth and uHealth* 11, e45464. ISSN: 2291-5222.

Fathi, Alireza, Xiaofeng Ren, and James M. Rehg (June 2011). "Learning to recognize objects in egocentric activities". In: *CVPR 2011*. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, CO, USA: IEEE, pp. 3281–3288. ISBN: 978-1-4577-0394-2.

Feng, Yongchao et al. (Apr. 13, 2025). *Vision-Language Model for Object Detection and Segmentation: A Review and Evaluation*. arXiv: 2504.09480[cs].

Gao, Yixin et al. (2014). "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) : A Surgical Activity Dataset for Human Motion Modeling". In: *In Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*. MICCAI Workshop. Vol. 3.

Gemma Team et al. (2025). *Gemma 3 Technical Report*. Version Number: 1.

Gonzalez, Glebys T et al. (Jan. 25, 2021). "From the Dexterous Surgical Skill to the Battlefield—A Robotics Exploratory Study". In: *Military Medicine* 186 (Supplement_1), pp. 288–294. ISSN: 0026-4075, 1930-613X.

Goyal, Raghav et al. (2017). "The "something something" video database for learning and evaluating visual common sense". In: Publisher: arXiv Version Number: 2.

Gu, Chunhui et al. (2017). "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions". In: Publisher: arXiv Version Number: 4.

Gupta, Deepak, Kush Attal, and Dina Demner-Fushman (Mar. 22, 2023). "A dataset for medical instructional video classification and question answering". In: *Scientific Data* 10.1, p. 158. ISSN: 2052-4463.

Hamet, Pavel and Johanne Tremblay (Apr. 2017). "Artificial intelligence in medicine". In: *Metabolism* 69, S36–S40. ISSN: 00260495.

Huaulmé, Arnaud, Kanako Harada, et al. (Apr. 27, 2023). *PEg TRAnsfer Workflow recognition challenge report: Does multi-modal data improve recognition?* arXiv: 2202.05821[cs].

Huaulmé, Arnaud, Duygu Sarikaya, et al. (Nov. 2021). "MIcro-surgical anastomose workflow recognition challenge report". In: *Computer Methods and Programs in Biomedicine* 212, p. 106452. ISSN: 01692607.

Kay, Will et al. (2017). "The Kinetics Human Action Video Dataset". In: Publisher: arXiv Version Number: 1.

Kim, Wonjae, Bokyung Son, and Ildoo Kim (July 18, 2021). "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 5583–5594.

Kuehne, H. et al. (Nov. 2011). "HMDB: A large video database for human motion recognition". In: *2011 International Conference on Computer Vision*. 2011 IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, pp. 2556–2563. ISBN: 978-1-4577-1102-2 978-1-4577-1101-5 978-1-4577-1100-8.

Kuo, Weicheng et al. (2022). *F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models*. Version Number: 2.

Li, Junnan et al. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. Version Number: 2.

Li, Kunchang et al. (Nov. 17, 2022). *UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer*. arXiv: 2211.09552[cs].

Li, Yanghao et al. (Mar. 30, 2022). *MViTv2: Improved Multiscale Vision Transformers for Classification and Detection*. arXiv: 2112.01526[cs].

Li, Yin, Miao Liu, and James M. Rehg (Oct. 31, 2020). *In the Eye of the Beholder: Gaze and Actions in First Person Video*. arXiv: 2006.00626[cs].

Liu, Haotian, Chunyuan Li, Yuheng Li, et al. (May 15, 2024). *Improved Baselines with Visual Instruction Tuning*. arXiv: 2310.03744[cs].

Liu, Haotian, Chunyuan Li, Qingyang Wu, et al. (2023). *Visual Instruction Tuning*. Version Number: 2.

Liu, Xiaonan et al. (Apr. 2018). "Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease". In: *Translational Research* 194, pp. 56–67. ISSN: 19315244.

Liu, Ze et al. (June 2022). "Video Swin Transformer". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, pp. 3192–3201. ISBN: 978-1-66546-946-3.

Lukežič, Alan et al. (July 2018). "Discriminative Correlation Filter with Channel and Spatial Reliability". In: *International Journal of Computer Vision* 126.7, pp. 671–688. ISSN: 0920-5691, 1573-1405. arXiv: 1611.08461[cs].

Mackenzie, Colin F. et al. (Apr. 2021). "Enhanced Training Benefits of Video Recording Surgery With Automated Hand Motion Analysis". In: *World Journal of Surgery* 45.4, pp. 981–987. ISSN: 0364-2313, 1432-2323.

Mirchi, Nykan et al. (Feb. 27, 2020). "The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine". In:

PLOS ONE 15.2. Ed. by Paweł Pławiak, e0229596. ISSN: 1932-6203.

Nwoye, Chinedu Innocent et al. (May 2022). "Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos". In: *Medical Image Analysis* 78, p. 102433. ISSN: 13618415. arXiv: 2109.03223[cs].

Patocka, Catherine et al. (Jan. 2024). "The Impact of Just-in-Time Simulation Training for Healthcare Professionals on Learning and Performance Outcomes: A Systematic Review". In: *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare* 19.1, S32–S40. ISSN: 1559-713X, 1559-2332.

Pirsiavash, H. and D. Ramanan (June 2012). "Detecting activities of daily living in first-person camera views". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI: IEEE, pp. 2847–2854. ISBN: 978-1-4673-1228-8 978-1-4673-1226-4 978-1-4673-1227-1.

Schmidt, Adam et al. (2021). "Multi-view Surgical Video Action Detection via Mixed Global View Attention". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen De Bruijne et al. Vol. 12904. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 626–635. ISBN: 978-3-030-87201-4 978-3-030-87202-1.

Shackelford, Stacy A. et al. (Aug. 2021). "Case-control analysis of prehospital death and prolonged field care survival during recent US military combat operations". In: *Journal of Trauma and Acute Care Surgery* 91.2, S186–S193. ISSN: 2163-0763, 2163-0755.

Sigurdsson, Gunnar A. et al. (Apr. 25, 2018). *Actor and Observer: Joint Modeling of First and Third-Person Videos*. arXiv: 1804.09627[cs].

Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah (2012). "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild". In: Publisher: arXiv Version Number: 1.

Stewart, Thomas and P Bird (Feb. 3, 2022). "Health economic evaluation: cost-effective strategies in humanitarian and disaster relief medicine". In: *BMJ Military Health*, e001859. ISSN: 2633-3767, 2633-3775.

Tao, Lingling et al. (2012). "Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation". In: *Information Processing in Computer-Assisted Interventions*. Ed. by Purang Abolmaesumi et al. Red. by David Hutchison et al. Vol. 7330. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 167–177. ISBN: 978-3-642-30617-4 978-3-642-30618-1.

Tong, Zhan et al. (Oct. 18, 2022). *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*. arXiv: 2203.12602[cs].

Torre, Fernando de la et al. (Apr. 2008). "Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database". In: *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*.

Valderrama, Natalia et al. (2022). "Towards Holistic Surgical Scene Understanding". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang et al. Vol. 13437. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 442–452. ISBN: 978-3-031-16448-4 978-3-031-16449-1.

Vannaprathip, Narumol et al. (Apr. 2025). "SDMentor: A virtual reality-based intelligent tutoring system for surgical decision making in dentistry". In: *Artificial Intelligence in Medicine* 162, p. 103092. ISSN: 09333657.

Wachs, Juan P., Andrew W. Kirkpatrick, and Samuel A. Tisherman (July 13, 2021). "Procedural Telementoring in Rural, Underdeveloped, and Austere Settings: Origins, Present Challenges, and Future Perspectives". In: *Annual Review of Biomedical Engineering* 23.1, pp. 115–139. ISSN: 1523-9829, 1545-4274.

Xiao, Bin et al. (Nov. 10, 2023). *Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks*. arXiv: 2311.06242[cs].

Yuan, Lu et al. (Nov. 22, 2021). *Florence: A New Foundation Model for Computer Vision*. arXiv: 2111.11432[cs].

Zhao, Hang et al. (Sept. 4, 2019). *HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization*. arXiv: 1712.09374[cs].

Zhao, Yue et al. (2022). *Learning Video Representations from Large Language Models*. Version Number: 1.

Zhuo, Yupeng, Andrew W. Kirkpatrick, et al. (2025). "Overview of the Trauma THOMPSON Challenge at MICCAI 2023". In: *AI for Brain Lesion Detection and Trauma Video Action Recognition*. Ed. by Rina Bao et al. Vol. 14567. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 47–60. ISBN: 978-3-031-71625-6 978-3-031-71626-3.

Zhuo, Yupeng, Andrew W. Kirkpatrick, et al. (2025). "The Trauma THOMPSON Challenge Report MICCAI 2023". In: *AI for Brain Lesion Detection and Trauma Video Action Recognition*. Ed. by Rina Bao et al. Vol. 14567. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 61–71. ISBN: 978-3-031-71625-6 978-3-031-71626-3.