EgoEMS: A High-Fidelity Multimodal Egocentric Dataset for Cognitive Assistance in Emergency Medical Services

Keshara Weerasinghe[1], Xueren Ge[1], Tessa Heick[1], Lahiru Nuwan Wijayasingha[1], Anthony Cortez[2], Abhishek Satpathy[1], John Stankovic[1], Homa Alemzadeh[1]

[1]School of Engineering and Applied Science
[2]School of Medicine
{cjh9fw,zar8jw,vht2gm,lnw8px,aec3gp,cqa3ym,jas9f,ha4d}@virginia.edu

Abstract

Emergency Medical Services (EMS) are critical to patient survival in emergencies, but first responders often face intense cognitive demands in high-stakes situations. AI cognitive assistants, acting as virtual partners, have the potential to ease this burden by supporting real-time data collection and decision making. In pursuit of this vision, we introduce EgoEMS, the first end-to-end, high-fidelity, multimodal, multiperson dataset capturing over 20 hours of realistic, procedural EMS activities from an egocentric view in 233 simulated emergency scenarios performed by 62 participants, including 46 EMS professionals. Developed in collaboration with EMS experts and aligned with national standards, EgoEMS is captured using an open-source, low-cost, and replicable data collection system and is annotated with keysteps, timestamped audio transcripts with speaker diarization, action quality metrics, and bounding boxes with segmentation masks. Emphasizing realism, the dataset includes responder-patient interactions reflecting real-world emergency dynamics. We also present a suite of benchmarks for real-time multimodal keystep recognition and action quality estimation, essential for developing AI support tools for EMS. We hope EgoEMS inspires the research community to push the boundaries of intelligent EMS systems and ultimately contribute to improved patient outcomes.

Code & Dataset — https://uva-dsa.github.io/EgoEMS
Extended Version — https://arxiv.org/pdf/2506.15028

Introduction

Every year more than 28 million emergency medical incidents are responded to in the U.S. (National Association of State EMS Officials (NASEMSO) 2020). Upon arrival at an incident scene, Emergency Medical Services (EMS) personnel must rapidly assess the situation, process complex information about victims and the environment, and provide emergency care before transferring patients to the hospital. In these safety-critical scenarios, patient survival hinges on rapid and accurate decision making. However, EMS responders often face overwhelming physical, mental, and emotional demands, resulting in cognitive overload, burnout, and increased risk of errors (Sweller 2011; Crowe et al. 2018).

With the recent rise of embedded Artificial Intelligence (AI) and large language models (LLMs) along with the rapid advancements of Augmented Reality (AR) technologies, there is tremendous potential to develop Intelligent Cognitive Assistants (ICAs) that can act as virtual partners to enhance situational awareness, guide critical procedures, and support training (Preum et al. 2021). Yet, the medical and EMS domains remain significantly underserved due to a lack of large-scale, high-fidelity labeled datasets and major privacy and security challenges.

Recent works have proposed innovative support systems and technologies to aid first responders in high-stakes environments. One example is the development of ICAs for real-time diagnosis and treatment decision support (Preum et al. 2021; Jin et al. 2023; Weerasinghe et al. 2024a; Preum et al. 2019, 2018; Shu et al. 2019; Ge et al. 2024). These systems can also serve as virtual coaches for training, helping novice responders build expertise through real-time feedback. However, existing ICAs are typically developed using datasets with limited fidelity and single modalities (primarily speech), which fail to capture the procedural complexity and unpredictability of real-world EMS settings. To provide accurate predictions and effective support in practical scenarios, ICAs must perceive the environment through multimodal sensing and interpret multiple responder activities from a first-person perspective in real-time.

Advances in egocentric datasets (Yang et al. 2025; Liu et al. 2022; Bansal, Arora, and Jawahar 2022; Wang et al. 2024) encourage the development of personal AI assistants for daily life and procedural task automation. Unlike exocentric (third-person) recordings, egocentric perspectives align more naturally with wearable AI systems and are particularly effective at capturing fine-grained hand-object interactions that are often occluded or outside the field of view in external camera setups (Bansal, Arora, and Jawahar 2022). However, existing datasets mostly focus on routine daily activities, lack multimodal integration, and are not designed for high-stakes domains like emergency medicine, where actions are complex, time-critical, and performed by teams. Ego-Exo4D (Grauman et al. 2024) includes a limited number of medical procedures, such as Cardiopulmonary Resuscitation (CPR) and COVID-19 testing. Other datasets, like EgoSurgery (Fujii et al. 2024) and Trauma Thompson (Birch et al. 2023), although focused on health domains,

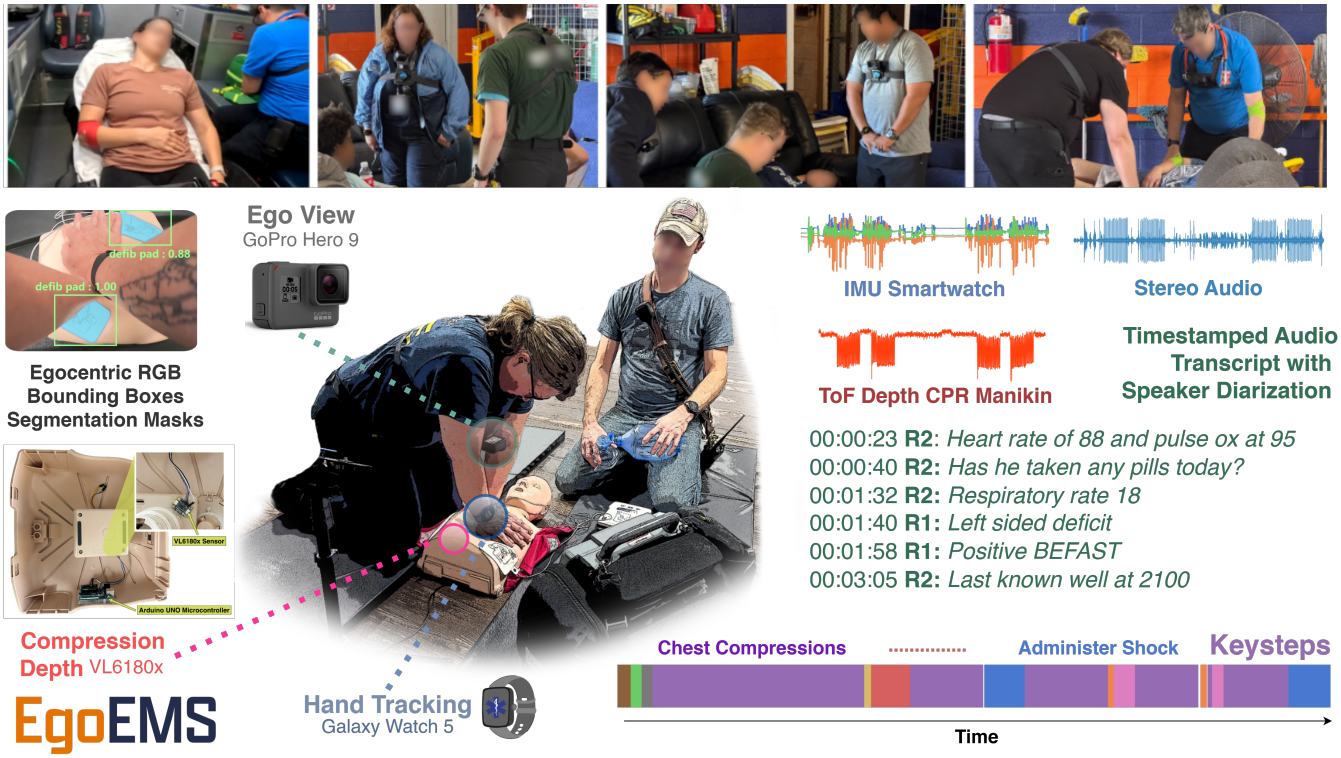

Figure 1: EgoEMS dataset provides synchronized egocentric multiperson views, along with rich high-fidelity multimodal data capturing EMS professionals in highly procedural tasks resulting in a total of 233 trials (20 hours) including 2694 keystep instances from commonly executed EMS interventions. The dataset includes annotations of keysteps, timestamped audio transcripts with speaker diarization, and semi-automatically generated bounding boxes and segmentation masks for key EMS objects, offering a comprehensive resource for understanding EMS workflows and developing AI solutions.

remain narrow in scope and limited to single views, specific procedures, and single modalities. Synthetic datasets, such as (Wang et al. 2023a), instead emphasize hand–tool pose estimation in controlled surgical scenes, but lack realism, procedural breadth, and real-world variability.

EMS incidents unfold in safety-critical environments where a team of first responders performs coordinated *keysteps* within standard procedures (*EMS interventions*) as defined by established *EMS protocol* guidelines under strict time constraints. These settings inherently generate rich multimodal information (e.g., egocentric views of medical interventions, conversational audio, and motion data) that can be leveraged for real-time decision support. For instance, Cardiac Arrest protocol requires maintaining correct CPR compression rate and depth for patient survival (Old Dominion EMS Alliance (ODEMSA) 2024; American Red Cross 2022) and proper timing of ventilations (e.g., 30 compressions to 2 breaths) for adequate oxygenation (see **Appendix A in Extended Version** for an overview of CPR procedure). Similarly, in stroke emergencies, rapid diagnosis and hospital transfer within clinically recommended time windows are required to minimize brain damage and improve outcomes (Saver 2006; Old Dominion EMS Alliance (ODEMSA) 2024). Although responders are trained on such guidelines, stress and cognitive load can impact their performance. ICAs can help track procedural keysteps and quality metrics, providing real-time reminders or feedback to improve adherence to protocols. However, the development of such systems is hindered by the absence of suitable datasets.

To address the aforementioned gaps, we introduce *EgoEMS, the first high-fidelity egocentric dataset designed specifically for cognitive assistance in EMS.* EgoEMS captures 20 hours of synchronized and labeled multimodal data of multiperson interactions in end-to-end EMS workflows from initial patient assessment to intervention, involving 62 subjects with varying skill levels (including EMS professionals and members of the public) performing 233 trials. Unlike prior datasets, EgoEMS is structured around nationally standardized EMS protocols (in the U.S.), spanning some of the most common EMS scenarios (cardiac arrest, cardiac suspected and stroke) and 9 critical interventions (including Airway-Breathing-Circulation (ABCs), 12-lead Electrocardiogram (ECG), CPR, Ventilation, Defibrillation, Stroke Assessment, Patient History, Vital Sign Assessment and Transport) according to NEMSIS (National EMS Information System 2024) database and expert feedback.

The dataset provides (i) egocentric views of the scene captured from responders' body-worn cameras, (ii) audio recordings of conversations at the scene, (iii) smartwatch IMU data capturing the responder's hand movements, and

(iv) ground-truth quality metrics (compression rate and depth during CPR procedures), synchronously collected using off-the-shelf components and custom open-source software (see Figure 1). We also provide annotations for EMS keysteps performed by responders, timestamped transcripts of responders' conversations with speaker diarization (i.e., automatic labeling of "who spoke when" in multi-speaker audio), and object bounding boxes with segmentation masks for the medical tools used by the responders in cardiac arrest emergencies. These annotations are generated using manual and semi-automatic approaches, based on the NREMT (National Registry of Emergency Medical Technicians 2024) guidelines and in collaboration with EMS experts (see Figure 2). Together, these elements enable end-to-end modeling of emergency response scenarios, from high-level protocol decisions to fine-grained action execution, with the hierarchical taxonomy providing structured representations of realistic EMS workflows.

We also present three benchmark tasks (see Figure 5), that reflect core real-time context inference capabilities essential for ICAs to support EMS responders: *Keystep Classification* for recognizing the specific keystep performed by a responder, *Keystep Segmentation* for detecting the start and end times of each keystep, and *Quality Evaluation* by continuous estimation of activity quality metrics (e.g., CPR compression rate and depth) utilizing multimodal data to provide feedback to responders.

In summary, our contributions are the following:

- *The first synchronized and labeled multimodal dataset of multiperson EMS procedural activities*, capturing collaborative dynamics of real-world scenarios with varied experience levels and certifications of EMS personnel.
- *A taxonomy of EMS activities, keysteps, and objects/-tools*, developed in collaboration with EMS professionals and aligned with NREMT, which is used to generate ground-truth annotations for activity recognition, and object detection and segmentation along with audio transcription and speaker diarization.
- *A suite of benchmarks for real-time activity recognition and quality estimation*, leveraging both single and combined modalities, to explore the performance of state-of-the-art (SOTA) supervised deep learning models compared to zero-shot methods including LLMs.
- *An open-source, low-cost, and easily replicable multimodal data collection system* based on off-the-shelf devices (e.g., GoPro Hero) and custom hardware integration (e.g., VL6180X ToF sensor) for synchronized capture of procedural activities through egocentric video and conversational audio recordings from the scene, smartwatch IMU data from hand movements, and ground-truth quality metrics from patient simulators.

Background and Related Work

Egocentric datasets. AI assistants that support real-world decision making require multimodal understanding of human activities from a first-person perspective (Preum et al. 2021). Existing egocentric datasets such as Epic-Kitchens (Damen et al. 2018), HOI4D (Liu et al. 2022), HoloAssist (Wang et al. 2023b), EgoVid-5M (Wang et al. 2024), EgoProceL (Bansal, Arora, and Jawahar 2022), and EgoLife (Yang et al. 2025) largely capture daily activities, object interactions, or scripted behaviors. Ego-Exo4D (Grauman et al. 2024) includes ego and exocentric views of limited medical procedures (e.g., CPR, COVID testing) performed by participants with basic training and accredited nurses. Other egocentric datasets in the medical domain include POV-Surgery (Wang et al. 2023a) and Ego-Surgery (Fujii et al. 2024) for open surgery and Trauma Thompson (Birch et al. 2023) for life-saving interventions such as tube thoracostomy and tracheostomy.

Despite these advances, none of the existing datasets capture the procedural structure, high-stakes environment, and coordinated multiperson nature of end-to-end EMS workflows (see Table 1). Curation of such datasets is particularly challenging due to privacy concerns and high annotation costs in medical settings, resulting in much smaller datasets. No prior dataset offers multimodal, multiperson egocentric recordings of simulated emergencies with certified responders and the ground truth necessary for modeling decision-making and cognitive assistance in critical care.

Activity recognition. Action recognition spans both classification, which assigns labels to video segments, and segmentation, which temporally localizes and labels actions over time. Prior work on egocentric video has addressed keystep or action classification by leveraging supervised deep learning models (Plizzari et al. 2023; Dessalene et al. 2023; Plizzari et al. 2022; Escorcia et al. 2022; Bansal, Arora, and Jawahar 2022; Grauman et al. 2024), with fewer efforts leveraging multimodal fusion of video and audio (Radevski et al. 2023; Gong et al. 2023). Segmentation has also been widely studied (Zhang, Wu, and Li 2022; Yi, Wen, and Jiang 2021; Li et al. 2020; Lea et al. 2017; Wang et al. 2016), though only a few works consider egocentric multimodal settings by combining video, audio, and IMU signals (Grauman et al. 2024; Huang et al. 2024).

In this paper, we take the first step towards benchmarking SOTA deep learning models for activity recognition using multimodal data for the EMS domain. We select strong, representative baselines that cover complementary approaches, including supervised models based on transformer architectures (Bertasius, Wang, and Torresani 2021; Weerasinghe et al. 2024b) capable of modeling long temporal dependencies and multimodal fusion, a convolutional TSN model (Wang et al. 2016) as a widely used CNN baseline, few-shot cross-domain models such as MM-CDFSL(Hatano et al. 2024) and zero-shot vision–language models such as Qwen-2.5 (Bai et al. 2025) and VideoLLaMA-3.3 (Zhang et al. 2025) to assess the potential of large pretrained models. For audio-only recognition, we also include a custom zero-shot pipeline using WhisperTimestamped (Louradour 2023) and GPT-4o (Achiam et al. 2023).

CPR quality estimation. CPR is one of the most safety-critical interventions in emergency care, where proper compression rate and depth are vital for patient survival (Ayala et al. 2014; Eftestøl et al. 2020). Although these metrics are central to effective feedback (Cheng et al. 2015b; Webber, Moran, and Cumin 2019; Cheng et al. 2015c), human

Dataset	Activity Setting	Synced	MP	IMU	Audio	RGB	Fine Act.	Tran-scripts	Obj. BB	Skill	Dur. (hrs)
Epic-Kitchens (Damen et al. 2018)	Daily Life	✗	✗	✗	✓	✓	✓	✓	✗	✗	100
HOI4D (Liu et al. 2022)	Object Manipulation	✗	✗	✗	✗	✓	✓	✗	✓	✗	44.4
EgoProceL (Bansal, Arora, and Jawahar 2022)	Daily Life	✗	✗	✗	✗	✓	✓	✗	✗	✗	62
HoloAssist (Wang et al. 2023b)	Daily Life	✓	✓	✓	✓	✓	✓	✓	✗	✓	166
EgoVid-5M (Wang et al. 2024)	Daily Life	✗	✗	✗	✗	✓	✓	✗	✗	✗	5550
EgoLife (Yang et al. 2025)	Daily Life	✓	✓	✗	✓	✓	✓	✓	✗	✗	300
Ego-Exo4D (Grauman et al. 2024)	Skilled Activities	✓	✗	✓	✓	✓	✓	✓	✓	✓	1442
Trauma-Thompson (Birch et al. 2023)	Medical Emergency	✗	✗	✗	✗	✓	✓	✗	✗	✗	∼1.6
EgoSurgery (Fujii et al. 2024)	Surgery	✗	✓	✗	✗	✓	✓	✗	✗	✗	15
POV-Surgery (Wang et al. 2023a)	Synthetic Surgery	✗	✗	✗	✗	✓	✗	✗	✓	✗	∼1
EgoEMS (Ours)	**Medical Emergency**	✓	✓	✓	✓	✓	✓	✓	✓	✓	**20**

Table 1: Comparison of egocentric datasets by modality availability and annotation types. Synced: Synchronized modalities, MP: Multiperson, Obj. BB: Object bounding boxes and segmentation masks, Fine Act: Fine-grained action annotations, Skill: Ground-truth for skill estimation, Dur: Approximate dataset duration (∼ estimated from reported frame counts at 30fps)

feedback is often biased (Jones et al. 2015). Recent work has explored skill assessment from egocentric video for general activities (Grauman et al. 2024; Huang et al. 2024), but CPR quality has largely been measured with accelerometer equipped CPRcards (Cheng et al. 2015a; Laerdal Medical 2024), defibrillator pads (González-Otero et al. 2015), or smartwatch IMU models sensitive to surface conditions (Lu et al. 2018; Jeong et al. 2015). Vision based methods using depth cameras (Loconsole et al. 2016; Di Mitri et al. 2019) require costly sensors in controlled, single-responder settings. In contrast, we introduce the first benchmark for quantitative, online CPR quality estimation that fuses egocentric video and smartwatch IMU to robustly predict compression rate and depth and enable real-time feedback in realistic, emergency scenarios.

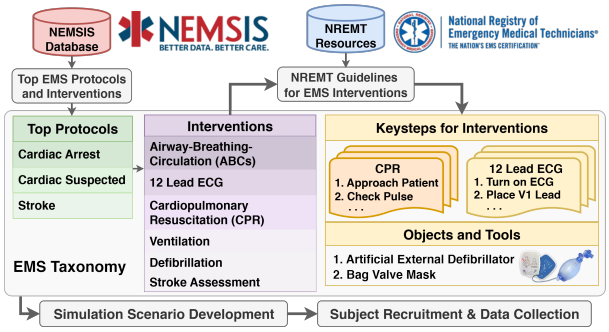

Figure 2: Methodology for creating the EgoEMS dataset including the EMS taxonomy.

EgoEMS Dataset

This section outlines the development of the EgoEMS dataset, including EMS taxonomy, simulation experiments, participants, data collection system and annotations. We refer the reader to Appendices A-C for more details.

EMS Taxonomy

To design a set of realistic EMS scenarios for data collection and adapt activity recognition models to EMS domain, we constructed a taxonomy of hierarchical EMS procedures, capturing high-level EMS protocols and their associated interventions and fine-grained keysteps (see Figure 2). First, we analyzed the NEMSIS database in consultation with EMS professionals to prioritize protocols that are high-frequency, time-sensitive and critical to patient survival. We focused on cardiac and stroke emergencies and further examined the distribution of interventions within those protocols to isolate most frequently performed interventions.

Specifically, we focus on scenarios involving the "Cardiac Arrest", "Chest-Pain Cardiac Suspected", "Stroke" protocols (Old Dominion EMS Alliance (ODEMSA) 2024) and 9 critical interventions associated with these protocols, including ABCs, 12-lead ECG, CPR, Ventilation, Defibrillation, Stroke Assessment, Patient History, Vital Signs Assessment and Transport. In collaboration with EMS experts, we define key procedural steps (keysteps) for selected interventions based on psychomotor examination guidelines from the NREMT, resulting in a total of 67 detailed keysteps representing the interventions. These keysteps are essential for detecting responders' actions within a protocol and evaluating proper procedural execution to provide continuous feedback. Finally, we identify the key EMS objects/tools used in Cardiac Arrest protocol as a part of the taxonomy, emphasizing the responder-equipment interactions that can improve EMS activity recognition. More details are in Appendix A.

Simulation Experiments

We conducted 233 simulated EMS trials spanning approximately 20 hours, encompassing a range of cardiac and stroke related emergency scenarios. The simulations were divided into two primary types: high-fidelity scenarios, where procedures were performed on human actors portraying patients, and cardiac arrest scenarios, which used manikins due to the nature of CPR intervention. High-fidelity simulations featured patients with diverse medical histories and demographics to promote data diversity and generalizabil-

Source	Subjects	Scenario	Interventions	Trials (Minutes)	Man. Ann.	Semi-auto Ann.			
					Keysteps	BB	SM	TT	RD
EMS Responders	46	Cardiac Arrest	CPR, Ventilation, Defibrillation	76 (183) ★	✓	✓	✓	✓	✓
		Cardiac Suspected	ABCs, 12-lead ECG	23 (173)	✓	✗	✗	✓	♢
		Stroke	ABCs, Stroke Assessment	41 (735)	✓	✗	✗	✓	♢
General Public	16	Cardiac Arrest	CPR	93 (116) ★	✓	♢	♢	♢	✓
Total	62			233 (1207)	2694	13.7k	12k	140	169

Table 2: Simulation scenarios, participants, and annotations in the dataset. ♢: Not applicable for the activity. ✗: Not provided due to limited visibility of the objects of interest. BB: Bounding boxes, SM: Segmentation masks, TT: Timestamped audio transcripts, RD: CPR compression rate and depth. ★: Trials that used manikins due to the nature of interventions and safety. The rest were high-fidelity with human patient actors.

ity. Each simulation captured end-to-end EMS procedures performed by a team of 2–3 responders, performing the critical interventions shown in Table 2. In addition, several scenarios were designed to reflect complex, realistic cases in which the initial chief complaint and presenting neurological symptoms mimicked a stroke but were ultimately attributable to alternative causes (e.g., hypoglycemia). These cases required responders to accurately assess, differentiate, and respond using appropriate protocol guided decision making. In cardiac arrest trials, EMS responders typically operated in pairs designated as *primary* and *secondary* responders carrying out critical interventions such as CPR, ventilation, and defibrillation on a manikin. In contrast, subjects from the general public conducted the trials individually and performed only CPR intervention on a manikin due to lack of medical training. To further enhance realism, additional volunteers served as bystanders, particularly in scenarios where the patient was unresponsive, contributing information such as patient history and situational context.

Participants A total of 62 participants were recruited, comprising 46 EMS professionals from 4 rescue squads and 16 individuals from the general public affiliated with an academic institution. EMS responders represented a broad range of experience levels and certifications, including members with basic CPR training, Emergency Medical Responders (EMRs), Emergency Medical Technicians (EMTs), and Paramedics. Years of experience ranged from under one year to over 30. The complexity of these simulations and the time required to perform them made large-scale data collection logistically challenging, as participating EMS agencies relied on volunteer personnel, many of whom were often called away for real emergency dispatches. Despite these challenges, the resulting dataset captures a wide range of realistic responder behaviors and skill levels. See Appendix B for more details.

Privacy and Ethics We obtained IRB approval prior to data collection, adhering to human subjects ethics. Real-world privacy and ethics considerations beyond simulation, as well as IRB and de-identification details are in the Ethical Statement and Appendix B.

Data Collection System

To capture EMS procedures, we used a remotely-controlled chest-mounted GoPro HERO camera to record the responder's egocentric view along with audio. The responder's dominant hand motions were tracked using a Samsung Galaxy Watch 5, recording 3-axis accelerometer data. Additionally, in cardiac arrest scenarios, chest compression ground truth metrics were measured using a VL6180X Time-of-Flight (ToF) sensor mounted on the manikin (see Figure 1). Synchronization was done based on Unix timestamps, with multiperson views synchronized along with all modalities, downsampled to align with GoPro's frame rate, resulting in a fully synchronized multimodal dataset. All data collection tools are available as open-source code, allowing others to replicate the system with low-cost and readily available hardware (see Appendix C).

Annotations

We employed manual and semi-automatic approaches, with some annotations purely manual and others using zero-shot models with manual verification. Table 2 shows a summary of annotations in the dataset. See Appendix B for details.

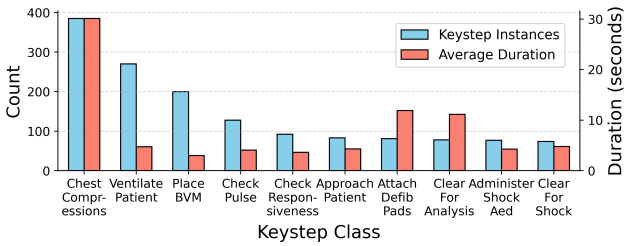

Figure 3: Top-10 keystep distribution with average duration. See Appendix A for a complete distribution.

Fine-grained keystep annotation. EgoEMS is manually annotated for 67 keysteps belonging to 9 interventions (see Appendix B and Figure 3). The dataset captures multiperson activity typical of real-world EMS procedures, where responders perform concurrent actions during the same time interval (e.g., AED activation during chest compressions; see Appendix A). While the dataset includes multiperson

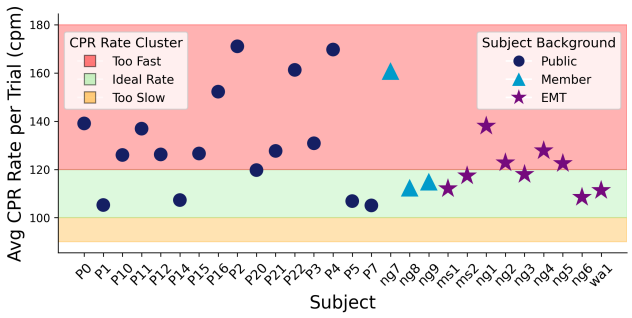

Figure 4: Ground-truth CPR compression rates for participants with varying skill levels. Each subject's reported rate is averaged across all their trials.

annotations, our benchmarks focus on the primary responder and their egocentric viewpoint. We do not leverage secondary responder annotations and multiview data for model training or evaluation, but include them to support future research on multiperson activity understanding.

Timestamped speaker diarization and transcription. We designed an automated LLM based pipeline to generate speaker-diarized timestamped transcripts for each trial's audio, which were then manually verified (see Appendix A). The general public participants struggled to narrate while performing CPR due to lack of advanced training, so they were instructed to focus solely on performing the task.

Bounding box and segmentation mask annotations. We generated bounding box and segmentation mask annotations for medical objects involved in cardiac arrest interventions (see Figure 1) using a semi-automatic pipeline. The nature of these interventions involves frequent and sustained interactions with critical medical equipment, making object localization particularly relevant in this context, serving as a promising candidate for improving the activity recognition capability of an ICA. This pipeline leverages a SOTA object detection model fine-tuned based on our EMS taxonomy combined with a zero-shot segmentation method. Our manual verification of 10% of the bounding box annotations against human annotations shows this method *saves over 60 hours of annotation time at a slight loss of precision.* We refer the reader to Appendix A.4.3 for more detailed analysis.

Compression depth and rate annotations. The ground-truth CPR depth and rate metrics were automatically generated by recording the compression depth using a ToF sensor integrated with a microcontroller embedded in the manikin (see Figure 1). Figure 4 shows the CPR rate distributions across skill levels where EMS professionals maintained steady CPR performance, while novice public participants showed much greater variability.

EgoEMS Benchmarks

We present three benchmark tasks designed to evaluate the core capabilities of an ICA for EMS: (1) *Keystep Classification* and (2) *Keystep Segmentation*, which together form the broader task of keystep recognition, a foundational capability for guiding responders through complex protocols while monitoring their actions and (3) *Action Quality Estimation*, which enables continuous feedback to improve execution quality. As illustrated in Figure 5, these tasks leverage synchronized multimodal data including egocentric video, audio, and smartwatch IMU to support *real-time inference* of procedural context and responder performance. Below, we provide an overview of each benchmark and results from a representative set of SOTA benchmarks. More detailed results and discussions are in Appendix D.

Benchmark 1: Keystep Classification

Motivation. Real-time keystep classification is a core capability for ICAs to guide responders through complex EMS protocols and monitor procedural adherence in real time. The intricacies of EMS interventions including rapid interactions with medical tools and parallel actions by multiple responders pose significant challenges for fine-grained action recognition. Egocentric video may suffer from occlusions or a limited field-of-view, while audio cues such as verbal requests for equipment can provide complementary information. Thus, drawing on insights from prior work (Yadav et al. 2021; Sun et al. 2023), leveraging multimodal data is essential for accurate activity classification.

Problem setting. We frame this as a multimodal action classification problem that aims to associate a data segment D_{seg} with a specific keystep in the set of keysteps in our EMS taxonomy. The trimmed data segments of a single modality or synchronized segments of multiple modalities, along with their associated keystep labels, are used during both training and testing. Given the scarcity of multimodal data in this domain, we also evaluate zero-shot methods, including LLMs, as baselines.

Summary results. We observed the highest top-1 accuracy of 62.3% using a supervised transformer model (Weerasinghe et al. 2024b) with egocentric video features extracted from a ResNet50 backbone (He et al. 2016), closely followed by 62.2% when smartwatch IMU and video data were fused together. While the fusion of these complementary modalities was expected to provide an improvement, the early fusion strategy we used did not yield additional gains, suggesting that more advanced fusion methods are needed to fully leverage long-range temporal and modality-specific cues. Notably, a zero-shot Qwen-2.5 (Bai et al. 2025) model achieves 38.3%, highlighting the potential of recent LLMs for activity recognition. See Appendix D.1 for detailed results and additional baselines.

Benchmark 2: Keystep Segmentation

Motivation. While classification assigns keysteps to fixed segments, ICA systems must also operate in online settings where actions unfold continuously. Keystep segmentation enables real-time tracking of procedural progress and timely intervention with a limited amount of context. However, achieving fine-grained segmentation is particularly challenging due to frequent view changes, variable execution speeds, and limited data per window. Multimodal sensing is essential for reliable performance in these dynamic settings.

Problem setting. We approach this as an online action segmentation task, aiming to identify and track specific keysteps performed by the primary responder throughout an

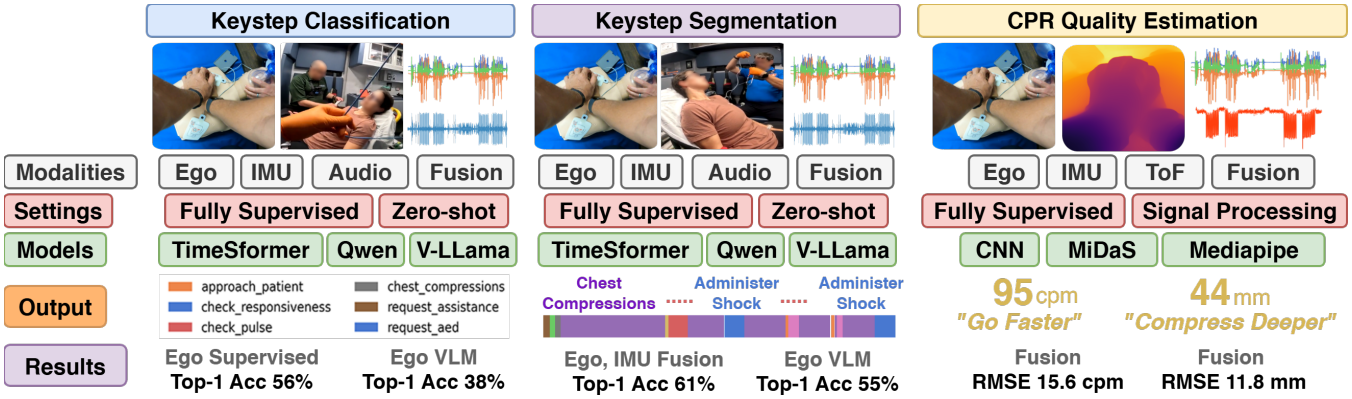

Figure 5: Overview of benchmark tasks for keystep classification, segmentation, and CPR quality estimation using multimodal data from EgoEMS. Example baseline models and representative results are shown to highlight the diverse evaluation scenarios enabled by the dataset.

EMS trial. Each trial is divided into 5-second streaming data segments, and each frame or sample within these windows is analyzed to classify the keystep occurring at that moment. Similar to keystep classification, zero-shot methods, including LLMs, are used as baselines.

Summary results. We achieved the best segmentation accuracy of 61% using a supervised transformer model (Weerasinghe et al. 2024b) with the fusion of egocentric video features and smartwatch IMU data. Unlike classification, this multimodal fusion provides a notable 6% improvement, likely attributable to the short temporal windows where complementary modalities help disambiguate subtle, fine-grained activities by better leveraging temporal cues. Interestingly, the zero-shot Qwen-2.5 (Bai et al. 2025) model achieved 55.5% accuracy, highlighting the potential of modern LLMs. WhisperTimestamped combined with GPT-4o (Achiam et al. 2023; Louradour 2023), using only audio, resulted in a lower accuracy of 38%, likely due to responders not consistently verbalizing their actions during critical interventions. See Appendix D.2 for a more detailed analysis.

Benchmark 3: CPR Quality Estimation

Motivation. An effective EMS ICA must be able to assess intervention quality to provide timely feedback. In CPR, compression rate and depth are critical quality metrics. The American Red Cross recommends 100–120 compressions per minute and a depth of at least 50 mm for adults (American Red Cross 2022). Deviations from these guidelines can compromise patient safety and reduce resuscitation success. We address this task by estimating compression rate and depth using egocentric video and smartwatch IMU data. The egocentric view captures close-up procedural context, while the repetitive hand motions in chest compressions make wrist IMU signals well-suited for dynamic estimation. By combining visual and inertial cues, an ICA can robustly infer CPR quality metrics in real time. We also propose a rule-based feedback generation framework that produces continuous, actionable insights to guide responder performance.

Problem setting. We formulate this task as an online recognition problem, where the model processes a 5-second sliding window of data, from either the chest-mounted GoPro or smartwatch IMU to estimate CPR quality metrics in real time. For each window, the model outputs the compression rate r (compressions per minute) and compression depth d (mm). The depth d is first computed for each individual compression within the window and then averaged to generate stable and actionable feedback. Ground-truth values from the ToF sensor are used for model supervision. Additionally, the rule-based feedback framework provides continuous feedback per window, which is also used to evaluate the model's performance (see Appendix D.3).

Summary results. Fusion of video and smartwatch IMU achieves the best overall CPR feedback performance, with an F1 score of 0.52 for compression rate and 0.83 for compression depth. While smartwatch IMU alone yields the lowest RMSE for compression rate, outperforming video and fusion, multimodal fusion provides the lowest RMSE for compression depth. See Appendix D.3 for more details.

Conclusion

EgoEMS is the first egocentric, multimodal, multiperson dataset dedicated to EMS, created to drive the development of AI systems that can function as virtual partners to first responders in the field. Developed in close collaboration with EMS professionals and aligned with national standards, it captures high-fidelity simulations of critical EMS procedures. EgoEMS provides over 20 hours of synchronized multimodal data, including 9 interventions and 67 keysteps across 233 trials, performed by 62 participants from multiple EMS agencies and the general public. The dataset's comprehensive taxonomy and fine-grained annotations are complemented by benchmark tasks that showcase the potential of cognitive assistants and support exploration of multimodal fusion. Furthermore, we provide open-source resources that enable reproducibility and future extensions. EgoEMS establishes a new benchmark for multimodal AI research and lays a robust foundation for next-generation EMS technologies that enhance responder performance, reduce cognitive burden, and improve patient outcomes.

Acknowledgments

This work was supported in part by the award 70NANB21H029 from the U.S. Department of Commerce, National Institute of Standards and Technology (NIST) and a research grant from the Commonwealth Cyber Initiative (CCI). We are grateful for the support and participation of several volunteer EMS first responders at rescue squads and fire agencies in the Charlottesville and Virginia Beach areas, specifically *North Garden Volunteer Fire Company*, *Western Albemarle Rescue Squad*, *Charlottesville Albemarle Rescue Squad* and *Ocean Park Volunteer Rescue Squad*, as well as George Stephens, James Fitz-Gerald, Margaret Sande, Jon Howard, and David Keeler, who helped us with organizing the simulation experiments.

Ethical Statement

This research was reviewed and approved by the Institutional Review Board for the Social and Behavioral Sciences (IRB-SBS), which established protocols for participant recruitment, informed consent, experimental procedures, and data management. All participants were fully informed of the study's purpose, procedures, potential risks and benefits, and future data usage through consent and materials release forms. Only data from participants who provided explicit consent for de-identified publication are included in the dataset. In accordance with the approved IRB protocol, all identifying information was removed prior to release: each participant was assigned a unique identifier, faces were blurred semi-automatically in egocentric video, and sensitive textual or audio identifiers such as names, license plates, and ID cards were manually obscured. All data underwent manual verification to ensure privacy preservation and compliance with ethical standards.

EgoEMS was developed using simulated emergency scenarios with trained responders under IRB oversight, enabling high-fidelity modeling of real-world conditions while maintaining participant safety and privacy. However, the extension of such systems to real-world EMS introduces substantial ethical and privacy challenges. Obtaining informed consent during emergencies is often infeasible when patients are unconscious or in critical conditions, and incidentally captured bystanders or minors may not have provided consent. Real scenes also encompass sensitive situations such as domestic violence, substance use, or mental health crises, which require enhanced safeguards, institutional oversight, and alignment with frameworks like HIPAA. Furthermore, ensuring robust de-identification and responsible data governance in unconstrained environments remains an open technical and ethical challenge.

Beyond data collection, the deployment of ICAs in EMS contexts raises broader societal considerations regarding surveillance, bias, and trust. Models trained on simulated data may underperform in the field due to shifts in environmental and behavioral distributions, necessitating rigorous validation, transparency, and accountability measures before clinical use. Our work therefore positions EgoEMS as a privacy conscious testbed to advance ICAs responsibly offering an open-source data collection framework, EMS-specific ontology, annotation tools, and de-identification pipeline to support gradual, ethically governed expansion to real-world settings.

We advocate for future efforts to establish consent and governance frameworks in collaboration with IRBs, EMS agencies, and legal experts; to develop policy-level standards inspired by body-worn camera protocols; and to implement on-device, HIPAA compliant data processing with built-in de-identification. At the algorithmic level, we emphasize the importance of multimodal learning methods that prioritize privacy-preserving modalities (e.g., IMU or depth sensors) when visual data are restricted. Together, these principles aim to ensure that cognitive assistance technologies for emergency response evolve safely, fairly, and in service of public trust and societal benefit.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

American Red Cross. 2022. CPR Steps.

Ayala, U.; Eftestøl, T.; Alonso, E.; Irusta, U.; Aramendi, E.; Wali, S.; and Kramer-Johansen, J. 2014. Automatic detection of chest compressions for the assessment of CPR-quality parameters. *Resuscitation*, 85(7): 957–963.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bansal, S.; Arora, C.; and Jawahar, C. 2022. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, 657–675. Springer.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Birch, E.; Couperus, K.; Gorbatkin, C.; Kirkpatrick, A. W.; Wachs, J.; Candelore, R.; Jiang, N.; Tran, O.; Beck, J.; Couperus, C.; et al. 2023. Trauma THOMPSON: clinical decision support for the frontline medic. *Military Medicine*, 188(Supplement_6): 208–214.

Cheng, A.; Brown, L. L.; Duff, J. P.; Davidson, J.; Overly, F.; Tofil, N. M.; Peterson, D. T.; White, M. L.; Bhanji, F.; and Bank, I. 2015a. Improving cardiopulmonary resuscitation with a CPR feedback device and refresher simulations (CPR CARES Study): a randomized clinical trial. *JAMA pediatrics*, 169(2): 137–144.

Cheng, A.; Hunt, E. A.; Grant, D.; Lin, Y.; Grant, V.; Duff, J. P.; White, M. L.; Peterson, D. T.; Zhong, J.; and Gottesman, R. 2015b. Variability in quality of chest compressions provided during simulated cardiac arrest across nine pediatric institutions. *Resuscitation*, 97: 13–19.

Cheng, A.; Overly, F.; Kessler, D.; Nadkarni, V. M.; Lin, Y.; Doan, Q.; Duff, J. P.; Tofil, N. M.; Bhanji, F.; and Adler, M. 2015c. Perception of CPR quality: influence of CPR feedback, just-in-time CPR training and provider role. *Resuscitation*, 87: 44–50.

Crowe, R. P.; Bower, J. K.; Cash, R. E.; Panchal, A. R.; Rodriguez, S. A.; and Olivo-Marston, S. E. 2018. Association of burnout with workforce-reducing factors among EMS professionals. *Prehospital Emergency Care*, 22(2): 229–236.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, 720–736.

Dessalene, E.; Maynord, M.; Fermüller, C.; and Aloimonos, Y. 2023. Therbligs in action: Video understanding through motion primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10618–10626.

Di Mitri, D.; Schneider, J.; Specht, M.; and Drachsler, H. 2019. Detecting mistakes in CPR training with multimodal data and neural networks. *Sensors*, 19(14): 3099.

Eftestøl, T.; Stokka, S. E.; Kvaløy, J. T.; Rad, A. B.; Irusta, U.; Aramendi, E.; Alonso, E.; Nordseth, T.; Skogvoll, E.; and Wik, L. 2020. A probabilistic function to model the relationship between quality of chest compressions and the physiological response for patients in cardiac arrest. In *2020 Computing in Cardiology*, 1–4. IEEE. ISBN 1728173825.

Escorcia, V.; Guerrero, R.; Zhu, X.; and Martinez, B. 2022. Sos! self-supervised learning over sets of handled objects in egocentric action recognition. In *European Conference on Computer Vision*, 604–620. Springer.

Fujii, R.; Hatano, M.; Saito, H.; and Kajita, H. 2024. EgoSurgery-Phase: a dataset of surgical phase recognition from egocentric open surgery videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 187–196. Springer.

Ge, X.; Satpathy, A.; Williams, R. D.; Stankovic, J.; and Alemzadeh, H. 2024. DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12798–12813. Miami, Florida, USA: Association for Computational Linguistics.

Gong, X.; Mohan, S.; Dhingra, N.; Bazin, J.-C.; Li, Y.; Wang, Z.; and Ranjan, R. 2023. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6481–6491.

González-Otero, D. M.; de Gauna, S. R.; Ruiz, J.; Daya, M. R.; Wik, L.; Russell, J. K.; Kramer-Johansen, J.; Eftestøl, T.; Alonso, E.; and Ayala, U. 2015. Chest compression rate feedback based on transthoracic impedance. *Resuscitation*, 93: 82–88.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Hatano, M.; Hachiuma, R.; Fujii, R.; and Saito, H. 2024. Multimodal Cross-Domain Few-Shot Learning for Egocentric Action Recognition. In *European Conference on Computer Vision (ECCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.; Chen, G.; Xu, J.; Zhang, M.; Yang, L.; Pei, B.; Zhang, H.; Dong, L.; Wang, Y.; Wang, L.; et al. 2024. EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-centric View of Procedural Activities in Real World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22072–22086.

Jeong, Y.; Chee, Y.; Song, Y.; and Koo, K. 2015. Smartwatch app as the chest compression depth feedback device. In *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*, 1465–1468. Springer.

Jin, L.; Liu, T.; Haroon, A.; Stoleru, R.; Middleton, M.; Zhu, Z.; and Chaspari, T. 2023. Emsassist: An end-to-end mobile voice assistant at the edge for emergency medical services. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 275–288.

Jones, A.; Lin, Y.; Nettel-Aguirre, A.; Gilfoyle, E.; and Cheng, A. 2015. Visual assessment of CPR quality during pediatric cardiac arrest: does point of view matter? *Resuscitation*, 90: 50–55.

Laerdal Medical. 2024. Get Familiar with CPRcard. https://laerdal.com/products/simulation-training/resuscitation-training/. Accessed: 2025-03-14.

Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 156–165.

Li, S.; Farha, Y. A.; Liu, Y.; Cheng, M.-M.; and Gall, J. 2020. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 6647–6658.

Liu, Y.; Liu, Y.; Jiang, C.; Lyu, K.; Wan, W.; Shen, H.; Liang, B.; Fu, Z.; Wang, H.; and Yi, L. 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21013–21022.

Loconsole, C.; Frisoli, A.; Semeraro, F.; Stroppa, F.; Mastronicola, N.; Filippeschi, A.; and Marchetti, L. 2016. RELIVE: a markerless assistant for CPR training. *IEEE Transactions on Human-Machine Systems*, 46(5): 755–760.

Louradour, J. 2023. whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.

Lu, T.-C.; Chen, Y.; Ho, T.-W.; Chang, Y.-T.; Lee, Y.-T.; Wang, Y.-S.; Chen, Y.-P.; Fu, C.-M.; Chiang, W.-C.; and Ma, M. H.-M. 2018. A novel depth estimation algorithm of chest compression for feedback of high-quality cardiopulmonary resuscitation based on a smartwatch. *Journal of biomedical informatics*, 87: 60–65.

National Association of State EMS Officials (NASEMSO). 2020. NASEMSO Releases 2020 National EMS Assessment. Accessed: 2024-11-09.

National EMS Information System. 2024. National EMS Information System (NEMSIS). https://nemsis.org/. Accessed: 2024-11-04.

National Registry of Emergency Medical Technicians. 2024. National Registry of Emergency Medical Technicians (NREMT). https://www.nremt.org/. Accessed: 2024-11-04.

Old Dominion EMS Alliance (ODEMSA). 2024. Regional Documents. Accessed: 2024-11-09.

Plizzari, C.; Perrett, T.; Caputo, B.; and Damen, D. 2023. What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13656–13666.

Plizzari, C.; Planamente, M.; Goletto, G.; Cannici, M.; Gusso, E.; Matteucci, M.; and Caputo, B. 2022. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19935–19947.

Preum, S.; Shu, S.; Hotaki, M.; Williams, R.; Stankovic, J.; and Alemzadeh, H. 2019. CognitiveEMS: A cognitive assistant system for emergency medical services. *ACM SIGBED Review*, 16(2): 51–60.

Preum, S. M.; Munir, S.; Ma, M.; Yasar, M. S.; Stone, D. J.; Williams, R.; Alemzadeh, H.; and Stankovic, J. A. 2021. A review of cognitive assistants for healthcare: Trends, prospects, and future directions. *ACM Computing Surveys (CSUR)*, 53(6): 1–37.

Preum, S. M.; Shu, S.; Ting, J.; Lin, V.; Williams, R.; Stankovic, J.; and Alemzadeh, H. 2018. Towards a cognitive assistant system for emergency response. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*, 347–348. IEEE.

Radevski, G.; Grujicic, D.; Blaschko, M.; Moens, M.-F.; and Tuytelaars, T. 2023. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5213–5224.

Saver, J. L. 2006. Time is brain—quantified. *Stroke*, 37(1): 263–266.

Shu, S.; Preum, S.; Pitchford, H. M.; Williams, R. D.; Stankovic, J.; and Alemzadeh, H. 2019. A behavior tree cognitive assistant system for emergency medical services. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6188–6195. IEEE.

Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; and Liu, J. 2023. Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3200–3225.

Sweller, J. 2011. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, 37–76. Elsevier.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36. Springer.

Wang, R.; Ktistakis, S.; Zhang, S.; Meboldt, M.; and Lohmeyer, Q. 2023a. POV-Surgery: A Dataset for Egocentric Hand and Tool Pose Estimation During Surgical Activities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 440–450.

Wang, X.; Kwon, T.; Rad, M.; Pan, B.; Chakraborty, I.; Andrist, S.; Bohus, D.; Feniello, A.; Tekin, B.; Frujeri, F. V.; et al. 2023b. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20270–20281.

Wang, X.; Zhao, K.; Liu, F.; Wang, J.; Zhao, G.; Bao, X.; Zhu, Z.; Zhang, Y.; and Wang, X. 2024. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*.

Webber, J.; Moran, K.; and Cumin, D. 2019. Paediatric cardiopulmonary resuscitation: Knowledge and perceptions of surf lifeguards. *Journal of Paediatrics and Child Health*, 55(2): 156–161.

Weerasinghe, K.; Janapati, S.; Ge, X.; Kim, S.; Iyer, S.; Stankovic, J. A.; and Alemzadeh, H. 2024a. Real-Time Multimodal Cognitive Assistant for Emergency Medical Services. *arXiv preprint arXiv:2403.06734*.

Weerasinghe, K.; Roodabeh, S. H. R.; Hutchinson, K.; and Alemzadeh, H. 2024b. Multimodal Transformers for Real-Time Surgical Activity Prediction. *arXiv preprint arXiv:2403.06705*.

Yadav, S. K.; Tiwari, K.; Pandey, H. M.; and Akbar, S. A. 2021. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223: 106970.

Yang, J.; Liu, S.; Guo, H.; Dong, Y.; Zhang, X.; Zhang, S.; Wang, P.; Zhou, Z.; Xie, B.; Wang, Z.; et al. 2025. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28885–28900.

Yi, F.; Wen, H.; and Jiang, T. 2021. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.