Trauma THOMPSON: A Dataset and Realistic Generative Framework for AI Copilots in Emergency Care

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

[1] Purdue University, West Lafayette, IN, USA
[2] University of Calgary, Calgary, Alberta, Canada
jpwachs@purdue.edu

Abstract

We introduce Trauma THOMPSON, a dataset and suite of benchmarks designed to accelerate the development of AI-powered copilots for real-time decision-making in emergency and resource-limited medical settings. This work proposes a method to address a critical bottleneck for future deployment: models trained on simulations may not perform well in the real world. The dataset features 3,717 unscripted, first-person video clips of five emergency procedures, uniquely including "just-in-time" (JIT) interventions that mirror the improvisational nature of field medicine. To obtain realistic patient data without ethical issues and identity concerns that medical data often encounter, we also propose TraumaGen, a novel framework for generating photorealistic patient and wound images from manikins while preserving clinical context. We establish benchmarks for action recognition, anticipation, and visual question answering (VQA), evaluating state-of-the-art models to demonstrate the challenges and potential of our dataset. By focusing on realism and improvisation, Trauma THOMPSON provides a crucial resource and a clear path toward developing and validating robust AI assistants for future deployment in real-world emergency care.

Datasets — https://doi.org/10.7910/DVN/V5BTRU

Introduction

Providing high-quality medical care in remote, disaster-stricken, and combat environments is challenging due to limited expertise, scarce resources, and unreliable connectivity (Wachs, Kirkpatrick, and Tisherman 2021). First responders often have minimal training yet must manage complex cases with constrained resources, increasing the risk of poor outcomes (Stewart and Bird 2022; Shackelford et al. 2021). To support medics during real-time treatment, AI copilots leveraging advances in generative models have been proposed (Bahl 2020; 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).

Most AI assistants are designed for operating rooms (OR) (Novaes and Basu 2020) or tested under laboratory conditions (Rojas, Couperus, and Wachs 2020; Xu, Islam, and Ren 2022), but these assumptions rarely hold in field medicine. When standard tools are unavailable, daily objects may be repurposed for care—for instance, tearing clothing to stop bleeding. While "just-in-time" (JIT) training improves outcomes (Patocka et al. 2024; Branzetti et al. 2017), little is known about training AI assistants for such conditions. Extending copilots beyond the OR into emergency and resource-limited environments remains a challenge.

A major gap is that current models cannot interpret visually complex settings or recognize improvised tools, such as belts as tourniquets or scissors as scalpels. Advanced computer vision could help AI identify injuries, recognize common objects, and suggest their use as JIT medical tools. The "simulation-to-reality" gap further limits progress: due to scarce patient data, most models are trained on manikins or simulators that lack diversity in anatomy, skin tone, and wounds, creating uncertainty about real-world safety.

To address these gaps, we introduce the Trauma THOMPSON Dataset (TTD), a collection of annotated video clips for developing AI copilots in low-resource and emergency care. To our knowledge, TTD is the first dataset of this scale and scope for field medicine. We also present TraumaGen, a method for realistic trauma scene generation. Figure 1 illustrates our experimental pipeline. In summary, this paper makes the following contributions.

1. Trauma THOMPSON Dataset (TTD): We introduce the first egocentric dataset for operational medicine, capturing field medics performing life-saving procedures in resource-constrained environments. The dataset covers both standard procedures with medical tools and JIT procedures using improvised objects, enabling rigorous testing of domain generalization.

2. Benchmarks for AI Copilots: TTD provides benchmarks for action recognition, anticipation, hand tracking, object detection, and visual question answering (VQA). In particular, VQA benchmarks highlight the potential for clinical decision support through natural dialogue.

3. Trauma Scene Generation (TraumaGen): We propose a method to generate photorealistic trauma scenes of real patients from manikin data, helping bridge the simulation-to-reality gap and improving the applicability of AI systems in real-world emergency care.

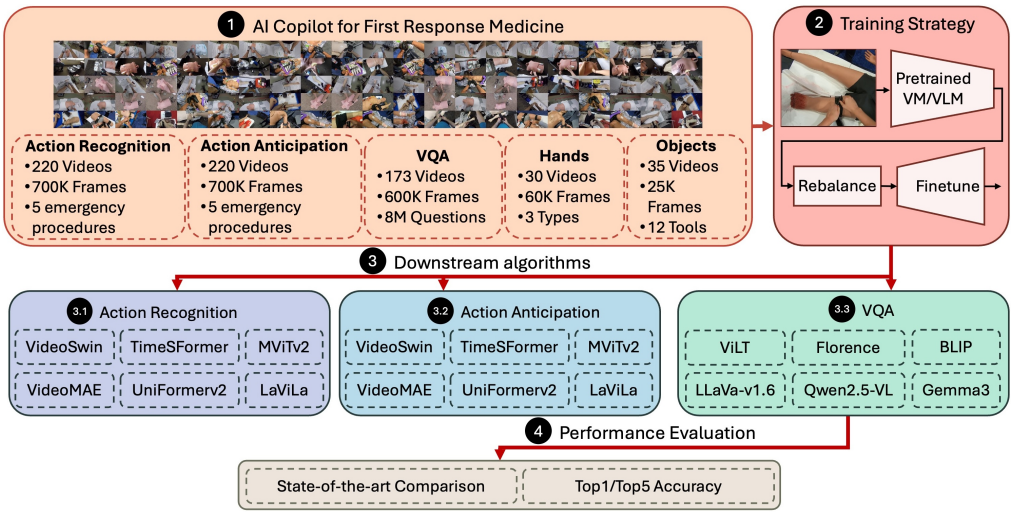


Figure 1: Overview of the experimental pipeline.

Related Work

Egocentric Activity Recognition and Surgical Datasets

Video understanding has advanced through 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. 2018), which mainly consist of short, scripted clips of single actions. To capture more realistic activities, research shifted toward first-person vision. Early efforts include daily living datasets (Pirsiavash and Ramanan 2012), cooking-focused EGTEA Gaze+ (Li, Liu, and Rehg 2020), and EPIC-KITCHENS, a large-scale egocentric cooking dataset (Damen et al. 2018).

In surgery, datasets have been developed for proficiency and skill evaluation (Gao et al. 2014; Tao et al. 2012; Gonzalez et al. 2021), while instructional videos for life-saving skills were suggested to train AI (Gupta, Attal, and Demner-Fushman 2023). However, most were collected in controlled environments using planned procedures or simulations.

Action Recognition and Anticipation

Human actions have been represented using visual (RGB, skeleton, depth, event streams) and non-visual (audio, motion, radar, WiFi) modalities. Despite challenges from viewpoint and illumination changes, RGB video remains the dominant modality (Sun et al. 2023). Deep learning approaches span CNNs, RNNs, and vision transformers (ViT) (Yang et al. 2022; Ulhaq et al. 2022), with ViTs gaining prominence due to attention mechanisms. Beyond recognition, anticipating future actions is critical for AI copilots, including early recognition, next-action prediction, and long-term anticipation (Roy, Rajendiran, and Fernando 2023). Recent work explores multimodal methods and large language models (LLMs) (Kong and Fu 2022; Yan et al. 2024; Xiang et al. 2023; Dessalene et al. 2023).

Medical VQA Datasets

Medical VQA has been studied through datasets such as VQA-MED (Hasan et al. 2018; Abacha et al. 2019, 2020; Ben Abacha et al. 2021), PathVQA (He et al. 2020), VQA-RAD (Lau et al. 2018), RadVisDial (Kovaleva et al. 2020), and SLAKE (Liu et al. 2021), which focus on medical images (CT, MRI, x-ray, ultrasound). VQA shows promise as a CDS tool by mimicking sequential diagnostic reasoning (Kim et al. 2024; Xu et al. 2024). However, progress is limited by scarce, diverse data (Antoniadi et al. 2021), especially in resource-limited settings where VQA could significantly enhance life-saving care.

Figure 2: Examples of JIT video clips.

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 emergency procedure 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+ (Li, 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

Dataset	Ego	Med	Austere	Frames	No. Act	Participants	No. Envs
TTD, 2025	✓	✓	✓	0.7M	162	12	15
EgoExOR, 2025 (Özsoy et al. 2025)	✓	✓	×	85K	22 Verb + 36 Noun	12	2
EgoSurgery-Phase, 2024 (Fujii et al. 2024)	✓	✓	×	28K	9	8	10
EvIs-Kitchen, 2024 (Hao et al. 2024)	✓	×	×	4527 Videos	35 Verb + 56 Noun	12	12
EPIC-KITCHENS, 2018 (Damen et al. 2018)	✓	×	×	11.5M	149	32	32
Ego-SLD, 2025 (Ali et al. 2025)	✓	×	×	892 videos	16	12	12
BEOID, 2014 (Damen 2014)	✓	×	×	0.1M	34	5	1
GTEA, 2011 (Fathi, Ren, and Rehg 2011)	✓	×	×	0.4M	42	13	1
CMU-MMAC, 2008 (de la Torre et al. 2008)	✓	×	×	0.2M	31	16	1
ADL, 2012 (Pirsiavash and Ramanan 2012)	✓	×	×	1.0M	32	20	20
ESAD, 2020 (Bawa et al. 2020)	×	✓	×	0.03M	21	4	4
CholecT50, 2022 (Nwoye et al. 2022)	×	✓	×	0.1M	100	13	13
MedVidCL, 2023 (Gupta, Attal, and Demner-Fushman 2023)	×	✓	×	1489 Videos	0	>100	>100
MRAO, 2021 (Schmidt et al. 2021)	×	✓	×	480 Videos	10	16	2
MISAW, 2021 (Huaulmé et al. 2021)	×	✓	×	27 Videos	17	6	1
PSI-AVA, 2022 (Valderrama et al. 2022)	×	✓	×	8 Videos	167	3	1
PETRAW, 2023 (Huaulmé et al. 2023)	×	✓	×	150 Videos	6	4	2

Table 1: Comparison of TTD to other egocentric and medical datasets. "Ego" denotes egocentric dataset. "Med" denotes medical dataset. "Austere" denotes low resource environments. "No. Act" denotes number of action classes. "No. Envs" denotes number of environment settings.

"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 highly variable style and performance time. Lastly, as opposed to existing datasets for surgical guidance and instruction, which rely on a standard medical tools, our dataset is subject to emergency settings. The performers were first responders, medics, and surgeons, and the procedures were conducted with improvised tools (e.g., using a shirt as a tourniquet or a pocket knife for cricothyroidotomy) to replicate a setting with limited resources, as shown in Figure 2. This presents a challenge for algorithms that rely on object detectors for activity recognition because the objects are unknown in emergent settings.

Dataset

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.

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 there is 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 acc $= \frac{\sum_{i=1}^{n}(p_i*(t_{ae}-t_{as}))}{\sum_{i=1}^{n}(t_{ae}-t_{as})}$.

Just-In-Time Procedures

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

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 VQA 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.*

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.

Dataset Class Distributions

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.

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.

Realistic Frame Generation

Stability Metrics

To assess the robustness of our frame generation framework and obtain best generated frame, we introduce two stability metrics:

- **Realism Stability (S_R)**: Measures the consistency of the realism classifier's outputs across multiple generations:

$$S_R = 1 - \frac{1}{N}\sum_{i=1}^{N} |r^{(i)} - \bar{r}|, \quad \text{where } \bar{r} = \frac{1}{N}\sum_{i=1}^{N} r^{(i)} \quad (1)$$

  where $r^{(i)} \in [0,1]$ is the realism score for the i-th generated frame, and N is the batch size.

- **Semantic Stability (S_C)**: Quantifies the preservation of clinical context during iterative refinements:

$$S_C = 1 - \frac{1}{N}\sum_{i=1}^{N} (\text{sim}^{(i)} - \mu)^2, \quad \mu = \frac{1}{N}\sum_{i=1}^{N} \text{sim}^{(i)} \quad (2)$$

  where $\text{sim}^{(i)} = \cos(\phi(\tilde{C}_S), \phi(\tilde{C}_R^{(i)}))$ is the caption similarity for the i-th generation.

- **Composite Stability Score (CSS)** The overall stability combines both metrics through a weighted average:

$$S = \lambda S_R + (1-\lambda) S_C, \quad \lambda \in [0,1] \quad (3)$$

  where λ controls the relative importance of realism versus semantic preservation.

Algorithm 1: TraumaGen – Stability Aware Realistic Frame Generation

Require: Source image I_S, VLM, generators G_H, G_W, masking function $\mathcal{M}$, realism metric $\mathcal{R}$, stability weight λ, thresholds δ, τ, τ_S

Ensure: Optimal realistic output image I^* or failure message

$C_S \leftarrow \text{VLM}(I_S)$ ▷ Generate initial caption
$\tilde{C}_S \leftarrow \mathcal{M}(C_S)$ ▷ Mask subject-specific terms
StabilityTracker $\leftarrow \emptyset$ ▷ Store $(r^{(i)}, \text{sim}^{(i)})$
repeat
  $\text{max\_score} \leftarrow -\infty$
  for $i \leftarrow 1$ to N **do**
    $I_H^{(i)} \leftarrow G_H(I_S, \tilde{C}_S)$ ▷ Generate human patient
    $I_R^{(i)} \leftarrow G_W(I_H^{(i)}, \tilde{C}_S)$ ▷ Generate wound
    $C_R^{(i)} \leftarrow \text{VLM}(I_R^{(i)})$, $\tilde{C}_R^{(i)} \leftarrow \mathcal{M}(C_R^{(i)})$
    $\text{sim}^{(i)} \leftarrow \cos(\phi(\tilde{C}_S), \phi(\tilde{C}_R^{(i)}))$
    $r^{(i)} \leftarrow \mathcal{R}(I_R^{(i)})$ ▷ Realism score
    $\text{score}^{(i)} \leftarrow \lambda r^{(i)} + (1-\lambda)\text{sim}^{(i)}$ ▷ Stability-aware scoring
    if $\text{score}^{(i)} > \text{max\_score}$ **then**
      $\text{max\_score} \leftarrow \text{score}^{(i)}$
      $I^* \leftarrow I_R^{(i)}$ ▷ Track current best image
    end if
    StabilityTracker.append$((r^{(i)}, \text{sim}^{(i)}))$
  end for
  $\bar{r} \leftarrow \text{mean}(\{r^{(i)}\}_{i=1}^N)$, $\bar{\text{sim}} \leftarrow \text{mean}(\{\text{sim}^{(i)}\}_{i=1}^N)$ ▷ Compute batch stability
  $S_R \leftarrow 1 - \frac{1}{N}\sum_{i=1}^{N} |r^{(i)} - \bar{r}|$ ▷ Realism stability
  $S_C \leftarrow 1 - \frac{1}{N}\sum_{i=1}^{N} (\text{sim}^{(i)} - \bar{\text{sim}})^2$ ▷ Semantic stability
  $S \leftarrow \lambda S_R + (1-\lambda) S_C$ ▷ Composite stability
  if $S < \tau_S$ **then**
    $\Delta_C \leftarrow \text{LM}(\tilde{C}_S, \text{mode}(\{\tilde{C}_R^{(i)}\}_{i=1}^N))$ ▷ Modal semantic drift
    $\tilde{C}_S \leftarrow \text{RefinePrompt}(\tilde{C}_S, \Delta_C)$
  end if
until $S \geq \tau_S$ ▷ Stability threshold
if $\mathcal{R}(I^*) \geq \tau$ **and** $S \geq \tau_S$ **then**
  return I^* ▷ Return stable and realistic image
else
  return "Failure: Unstable or unrealistic generation"
end if

Algorithm for Realistic Frame Generation

Our stability-aware generation framework transforms a source image (I_S) of a manikin into a photorealistic patient (I_R) while preserving clinical context through a multi-stage process, as shown in Algorithm 1. Initially, a vision-language model (VLM) generates a caption C_S for I_S, which is processed by a masking function $\mathcal{M}$ to produce a subject-neutral prompt $\tilde{C}_S$. The system then generates a batch of N candidate images through component-wise synthesis: first applying human patient generation G_H, followed by wound generation G_W. For each candidate $I_R^{(i)}$,

we compute both realism scores $r^{(i)} = \mathcal{R}(I_R^{(i)})$ and semantic similarity $\text{sim}^{(i)} = \cos(\phi(\tilde{C}_S), \phi(\tilde{C}_R^{(i)}))$ where $\tilde{C}_R^{(i)} = \mathcal{M}(\text{VLM}(I_R^{(i)}))$.

The optimal image I^* is selected by maximizing the stability-weighted score $\lambda r^{(i)} + (1-\lambda)\text{sim}^{(i)}$, where $\lambda \in [0,1]$ controls the realism-semantics trade-off. Batch stability is evaluated through:

$$S = \lambda \left(1 - \frac{1}{N}\sum_{i=1}^{N} |r^{(i)} - \bar{r}|\right) + (1-\lambda)\left(1 - \frac{1}{N}\sum_{i=1}^{N}(\text{sim}^{(i)} - \mu)^2\right)$$

with $\bar{r}$ and μ being batch means. The generation iterates until $S \geq \tau_S$ (stability threshold) and $\text{sim}^{(i)} \geq \delta$, refining $\tilde{C}_S$ using semantic drift $\Delta_C = \text{LM}(\tilde{C}_S, \text{mode}(\{\tilde{C}_R^{(i)}\}))$ when needed. Implementations used ChatGPT-4o (VLM), Flux.1-Kontext (G_H, G_W), and Qwen2.5 (LM). For all experiments, we set $\lambda = 0.7$ (favoring realism), $\tau_S = 0.85$ (stability threshold), $\delta = 0.95$ (semantic similarity threshold), and $N = 3$ (batch size), with $\mathcal{R}$ requiring $r^{(i)} \geq 0.95$ for photorealistic acceptance. For realism assessment, we implemented a ResNet50 classifier $\mathcal{R}$ to differentiate real patients from manikins, achieving 100.0% accuracy on our hold-out test set.

Figure 3: Realistic images generated through our TraumaGen framework. Top row images are manikin/simulation and the bottom row corresponds to generated images.

Realistic Frames

Our TraumaGen framework transforms clinical training images featuring manikins into photorealistic outputs with human patients as shown in Figure 3. The results demonstrate our model's ability to consistently generate medically relevant details, such as complex wounds and active bleeding. A key strength of our approach is the preservation of the original context, such as background elements, medical tools, and personnel, ensuring the clinical scenario is accurately maintained. This highlights the controllability of our method, which allows us to produce graphic yet necessary medical content without triggering the content violation issues commonly encountered with proprietary models.

Benchmark Results

Action Recognition and Anticipation

Evaluation Setups and Metrics All models were trained using either only the regular data or a combination of both regular and JIT frames. They were evaluated under three conditions: only regular procedures, only JIT procedures, and the combination of both. For both standard and JIT video operations, the dataset was split into training and testing sets, with 80% of the videos used for training and 20% for testing. Training was conducted using the RTX™ 4090 Ti GPU. A class-agnostic evaluation method, following the approach in (Zhao et al. 2019), was used to measure model accuracy. The models' performance was evaluated using Top-1 and Top-5 accuracy metrics for verbs, nouns, and full actions (a combination of verb and noun).

Performance We trained both vision models (VMs) and vision-language models (VLMs) for action recognition and anticipation on regular and JIT procedures of the TTD dataset. Pretrained weights were fine-tuned on TTD, with random oversampling used for class balancing. Table 2 report Top-1 and Top-5 accuracy for six models across three testing settings.

For training on regular procedure and testing on regular procedure, MViT v2 achieves the highest accuracy in both recognition and anticipation, while TimeSFormer performs worst. Performance drops sharply when models trained on regular data are tested on JIT data: for recognition, MViT v2 reaches 14.49% Top-1, with TimeSFormer and LaViLa below 1%; for anticipation, MViT v2 and Uniformer v2 drop to 8% Top-1. This trend holds in the combined test set, indicating that regular data alone is insufficient for generalizing to JIT scenarios.

Training on combined data significantly improves JIT performance. For recognition, Uniformer v2 attains the best Top-1 accuracy, while MViT v2 leads in Top-5; for anticipation, MViT v2 and Uniformer v2 consistently outperform other models. On the combined test set, MViT v2 remains strongest, while TimeSFormer consistently ranks lowest. These results highlight the benefit of diverse training material for both recognition and anticipation tasks.

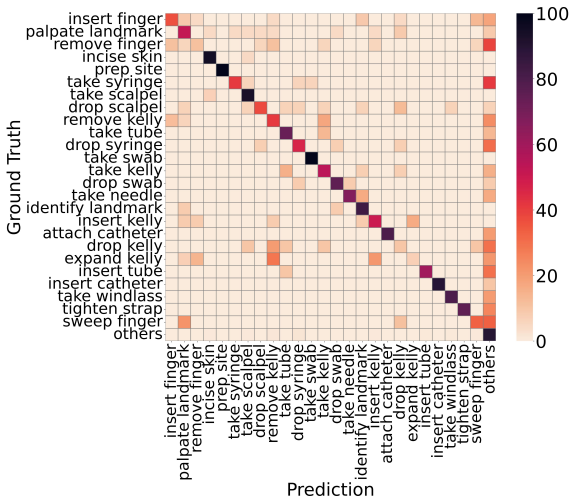
Overall, anticipation accuracy is lower than recognition, particularly under JIT testing, reflecting the added challenge of predicting future actions. Figures 4a–b illustrate MViT v2's confusion matrices, showing strong predictions on frequent classes but lower accuracy on less common actions.

VQA

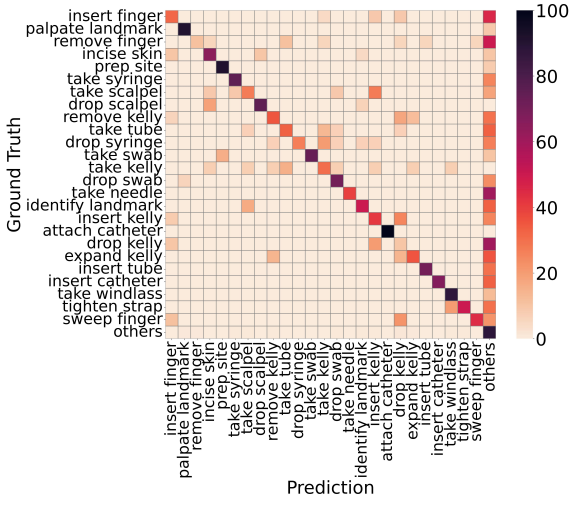
Table 3 compares the performance of 6 finetuned VQA models of various sizes, ViLT-B/32 (Kim, Son, and Kim 2021), BLIP (Li et al. 2022), Florence2(Yuan et al. 2021), LLaVA-v1.6-7B (Liu et al. 2023, 2024), Qwen2.5-VL-7B (Bai et al. 2025), and Gemma3-4B (Gemma Team et al. 2025). LLaVA-v1.6, Qwen2.5-VL and Gemma3 are finetuned with the QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al. 2023) approach for efficient adaptation on a single GPU. BLIP demonstrated strong VQA skills with a relatively moderate size, achieving the best accuracy of 88.64%. 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%. With 87 million parameters, ViLT-B/32 provides a lighter option.

Task	Train	Test	VideoSwin		TimeSFormer		VideoMAE		Uniformer v2		MViT v2		LaViLa	
			Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Recognition	Regular	Regular	45.10	74.52	31.91	62.81	43.34	71.89	60.47	85.65	65.59	89.75	42.52	68.81
		JIT	3.85	15.38	0.51	5.77	5.77	17.31	8.65	21.15	14.49	35.20	0.96	9.62
		Combined	34.60	66.71	27.70	55.27	38.37	64.68	53.62	77.13	58.58	82.08	38.25	60.99
	Combined	Regular	44.51	73.35	29.42	63.69	48.61	73.06	60.32	84.19	66.47	89.17	40.17	66.67
		JIT	39.42	70.19	32.69	58.65	44.23	65.38	53.85	80.77	50.96	90.38	37.71	63.84
		Combined	43.84	72.94	29.86	63.02	48.03	72.05	59.47	83.74	64.42	88.82	39.53	66.02
Anticipation	Regular	Regular	39.88	69.24	28.44	59.97	41.42	69.24	56.25	84.70	60.12	87.02	38.79	67.85
		JIT	3.16	7.37	3.16	7.37	2.11	8.42	5.26	23.16	8.42	22.11	1.05	8.42
		Combined	35.18	61.32	25.20	53.23	36.39	61.46	49.73	76.82	53.50	78.71	33.96	60.24
	Combined	Regular	41.89	69.09	26.74	61.21	44.05	69.40	57.95	83.77	60.74	86.56	40.03	67.58
		JIT	36.84	63.16	24.21	56.84	33.68	68.42	47.37	81.05	48.42	86.32	35.64	61.96
		Combined	41.24	68.33	26.42	60.65	42.72	69.27	56.60	83.42	59.16	86.52	39.65	66.73

Table 2: Performance comparison of action recognition and anticipation models on different train-test settings. Gray represents lower values, while blue represents higher values.



(a) Action recognition.



(b) Action anticipation.

Figure 4: Confusion matrices of action recognition and action anticipation with MViT v2 on regular procedures.

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

Table 3: VQA model performance comparison.

Conclusion

In this work, we present the first egocentric dataset of 5 emergency procedures performed in austere settings, introducing critical benchmarks for AI in operational medicine. Our dataset reflects real-world field situations where improvised equipment and varying levels of experience have a substantial impact on results. These issues include hand occlusions, procedural errors, and unscripted variances among performers. Furthermore, the inclusion of JIT procedures poses a novel challenge for VMs in zero-shot inference, emphasizing the need for robust generalization. Moreover, we propose a stability-aware generative framework that transforms manikin-based simulations into photorealistic patient scenarios. This method bridges the "sim-to-real" gap in AI medical training by enabling comprehensive testing of AI assistants trained on synthetic data. Models optimized for simulator environments often fail in real deployments due to distribution shifts and our approach mitigates this by generating clinically consistent and high-fidelity patients. Future work will include annotating the actions with whole sentences and detailed instructions, enlarging the dataset with more procedural settings, such as more spontaneous acts and larger contexts (e.g., more procedures), and generating improved realistic frames. In the intricate field of humanitarian medicine, such initiatives are essential for advancing the development of AI medical assistants. Moreover, training unified VLMs that can support all surgical tasks will greatly increase the clinical value of AI medical assistants.

Acknowledgements

This work was partially supported by the Center for AI and Robotic Excellence in medicine (CARE) at Purdue University and Indiana University School of Medicine. This work was also 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.

References

Abacha, A. B.; Datla, V. V.; Hasan, S. A.; Demner-Fushman, D.; and Müller, H. 2020. Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain. In *CLEF 2020 Working Notes*, CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org.

Abacha, A. B.; Hasan, S. A.; Datla, V. V.; Liu, J.; Demner-Fushman, D.; and Müller, H. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. *CLEF (working notes)*, 2(6).

Al-Antari, M. A. 2023. Artificial Intelligence for Medical Diagnostics—Existing and Future AI Technology! *Diagnostics*, 13(4): 688.

Ali, A.; Das, B.; Al-Shamayleh, A. S.; Sarkar, S.; Ray, C.; Mandal, S.; Sk, S.; and Akhunzada, A. 2025. Descriptor: Egocentric Action Recognition for Bengali Sign Language Detection Dataset (Ego-SLD). *IEEE Data Descriptions*, 2: 102–112. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Antoniadi, A. M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B. A.; and Mooney, C. 2021. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, 11(11): 5088.

Azari, D. P.; Frasier, L. L.; Quamme, S. R. P.; Greenberg, C. C.; Pugh, C. M.; Greenberg, J. A.; and Radwin, R. G. 2019. Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Annals of Surgery*, 269(3): 574–581.

Bahl, M. 2020. Artificial Intelligence: A Primer for Breast Imaging Radiologists. *Journal of Breast Imaging*, 2(4): 304–314.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. ArXiv:2502.13923 [cs].

Bawa, V. S.; Singh, G.; KapingA, F.; Skarga-Bandurova, I.; Leporini, A.; Landolfo, C.; Stabile, A.; Setti, F.; Muradore, R.; Oleari, E.; and Cuzzolin, F. 2020. ESAD: Endoscopic Surgeon Action Detection Dataset. ArXiv:2006.07164 [cs].

Ben Abacha, A.; Sarrouti, M.; Demner-Fushman, D.; Hasan, S. A.; and Müller, H. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

Branzetti, J. B.; Adedipe, A. A.; Gittinger, M. J.; Rosenman, E. D.; Brolliar, S.; Chipman, A. K.; Grand, J. A.; and Fernandez, R. 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. *BMJ Quality & Safety*, 26(11): 881–891.

Caballero, D.; Sánchez-Margallo, J. A.; Pérez-Salazar, M. J.; and Sánchez-Margallo, F. M. 2025. Applications of Artificial Intelligence in Minimally Invasive Surgery Training: A Scoping Review. *Surgeries*, 6(1): 7.

Damen, D. 2014. Bristol Egocentric Object Interactions Dataset.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. ArXiv:1804.02748 [cs].

de la Torre, F.; Hodgins, J.; Bargteil, A.; Collado, A.; Martin, X.; Macey, J.; and Beltran, P. 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*.

Dessalene, E.; Maynord, M.; Fermüller, C.; and Aloimonos, Y. 2023. LEAP: LLM-Generation of Egocentric Action Programs. *arXiv preprint arXiv:2312.00055*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs].

Dilsizian, S. E.; and Siegel, E. L. 2014. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. *Current Cardiology Reports*, 16(1): 441.

Dinh, A.; Yin, A. L.; Estrin, D.; Greenwald, P.; and Fortenko, A. 2023. Augmented Reality in Real-time Telemedicine and Telementoring: Scoping Review. *JMIR mHealth and uHealth*, 11: e45464.

Fathi, A.; Ren, X.; and Rehg, J. M. 2011. Learning to recognize objects in egocentric activities. In *CVPR 2011*, 3281–3288. Colorado Springs, CO, USA: IEEE. ISBN 978-1-4577-0394-2.

Feng, Y.; Liu, Y.; Yang, S.; Cai, W.; Zhang, J.; Zhan, Q.; Huang, Z.; Yan, H.; Wan, Q.; Liu, C.; Wang, J.; Lv, J.; Liu, Z.; Shi, T.; Liu, Q.; and Wang, Y. 2025. Vision-Language Model for Object Detection and Segmentation: A Review and Evaluation. ArXiv:2504.09480 [cs].

Fujii, R.; Hatano, M.; Saito, H.; and Kajita, H. 2024. EgoSurgery-Phase: A Dataset of Surgical Phase Recognition from Egocentric Open Surgery Videos. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.;

Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006, 187–196. Cham: Springer Nature Switzerland. ISBN 978-3-031-72088-8 978-3-031-72089-5. Series Title: Lecture Notes in Computer Science.

Gao, Y.; Vedula, S. S.; Reiley, C. E.; Ahmidi, N.; Varadarajan, B.; Lin, H. C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D. D.; Chen, C. C. G.; Vidal, R.; Khudanpur, S.; and Hager, G. 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)*, volume 3.

Gemma Team; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; Rouillard, L.; Mesnard, T.; Cideron, G.; Grill, J.-b.; Ramos, S.; Yvinec, E.; Casbon, M.; Pot, E.; Penchev, I.; Liu, G.; Visin, F.; Kenealy, K.; Beyer, L.; Zhai, X.; Tsitsulin, A.; Busa-Fekete, R.; Feng, A.; Sachdeva, N.; Coleman, B.; Gao, Y.; Mustafa, B.; Barr, I.; Parisotto, E.; Tian, D.; Eyal, M.; Cherry, C.; Peter, J.-T.; Sinopalnikov, D.; Bhupatiraju, S.; Agarwal, R.; Kazemi, M.; Malkin, D.; Kumar, R.; Vilar, D.; Brusilovsky, I.; Luo, J.; Steiner, A.; Friesen, A.; Sharma, A.; Sharma, A.; Gilady, A. M.; Goedeckemeyer, A.; Saade, A.; Kolesnikov, A.; Bendebury, A.; Abdagic, A.; Vadi, A.; György, A.; Pinto, A. S.; Das, A.; Bapna, A.; Miech, A.; Yang, A.; Paterson, A.; Shenoy, A.; Chakrabarti, A.; Piot, B.; Wu, B.; Shahriari, B.; Petrini, B.; Chen, C.; Lan, C. L.; Choquette-Choo, C. A.; Carey, C.; Brick, C.; Deutsch, D.; Eisenbud, D.; Cattle, D.; Cheng, D.; Paparas, D.; Sreepathihalli, D. S.; Reid, D.; Tran, D.; Zelle, D.; Noland, E.; Huizenga, E.; Kharitonov, E.; Liu, F.; Amirkhanyan, G.; Cameron, G.; Hashemi, H.; Klimczak-Plucińska, H.; Singh, H.; Mehta, H.; Lehri, H. T.; Hazimeh, H.; Ballantyne, I.; Szpektor, I.; Nardini, I.; Pouget-Abadie, J.; Chan, J.; Stanton, J.; Wieting, J.; Lai, J.; Orbay, J.; Fernandez, J.; Newlan, J.; Ji, J.-y.; Singh, J.; Black, K.; Yu, K.; Hui, K.; Vodrahalli, K.; Greff, K.; Qiu, L.; Valentine, M.; Coelho, M.; Ritter, M.; Hoffman, M.; Watson, M.; Chaturvedi, M.; Moynihan, M.; Ma, M.; Babar, N.; Noy, N.; Byrd, N.; Roy, N.; Momchev, N.; Chauhan, N.; Bunyan, O.; Botarda, P.; Caron, P.; Rubenstein, P. K.; Culliton, P.; Schmid, P.; Sessa, P. G.; Xu, P.; Stanczyk, P.; Tafti, P.; Shivanna, R.; Wu, R.; Pan, R.; Rokni, R.; Willoughby, R.; Vallu, R.; Mullins, R.; Jerome, S.; Smoot, S.; Girgin, S.; Iqbal, S.; Reddy, S.; Sheth, S.; Põder, S.; Bhatnagar, S.; Panyam, S. R.; Eiger, S.; Zhang, S.; Liu, T.; Yacovone, T.; Liechty, T.; Kalra, U.; Evci, U.; Misra, V.; Roseberry, V.; Feinberg, V.; Kolesnikov, V.; Han, W.; Kwon, W.; Chen, X.; Chow, Y.; Zhu, Y.; Wei, Z.; Egyed, Z.; Cotruta, V.; Giang, M.; Kirk, P.; Rao, A.; Lo, J.; Moreira, E.; Martins, L. G.; Sanseviero, O.; Gonzalez, L.; Gleicher, Z.; Warkentin, T.; Mirrokni, V.; Senter, E.; Collins, E.; Barral, J.; Ghahramani, Z.; Hadsell, R.; Matias, Y.; Sculley, D.; Petrov, S.; Fiedel, N.; Shazeer, N.; Vinyals, O.; Dean, J.; Hassabis, D.; Kavukcuoglu, K.; Farabet, C.; Buchatskaya, E.; Alayrac, J.-B.; Anil, R.; Dmitry; Lepikhin; Borgeaud, S.; Bachem, O.; Joulin, A.; Andreev, A.; Hardin, C.; Dadashi, R.; and Hussenot, L. 2025. Gemma 3 Technical Report. Version Number: 1.

Gonzalez, G. T.; Kaur, U.; Rahman, M.; Venkatesh, V.; Sanchez, N.; Hager, G.; Xue, Y.; Voyles, R.; and Wachs, J. 2021. From the Dexterous Surgical Skill to the Battlefield—A Robotics Exploratory Study. *Military Medicine*, 186(Supplement_1): 288–294.

Goyal, R.; Kahou, S. E.; Michalski, V.; Materzyńska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; Hoppe, F.; Thurau, C.; Bax, I.; and Memisevic, R. 2017. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*. Publisher: arXiv Version Number: 2.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6047–6056.

Gupta, D.; Attal, K.; and Demner-Fushman, D. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1): 158.

Hamet, P.; and Tremblay, J. 2017. Artificial intelligence in medicine. *Metabolism*, 69: S36–S40.

Hao, Y.; Kanezaki, A.; Sato, I.; Kawakami, R.; and Shinoda, K. 2024. Egocentric Human Activities Recognition With Multimodal Interaction Sensing. *IEEE Sensors Journal*, 24(5): 7085–7096. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Hasan, S. A.; Ling, Y.; Farri, O.; Liu, J.; Müller, H.; and Lungren, M. P. 2018. Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. In *Conference and Labs of the Evaluation Forum*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. PathVQA: 30000+ Questions for Medical Visual Question Answering. ArXiv:2003.10286 [cs].

Huaulmé, A.; Harada, K.; Nguyen, Q.-M.; Park, B.; Hong, S.; Choi, M.-K.; Peven, M.; Li, Y.; Long, Y.; Dou, Q.; Kumar, S.; Lalithkumar, S.; Hongliang, R.; Matsuzaki, H.; Ishikawa, Y.; Harai, Y.; Kondo, S.; Mitsuishi, M.; and Jannin, P. 2023. PEg TRAnsfer Workflow recognition challenge report: Does multi-modal data improve recognition? ArXiv:2202.05821 [cs].

Huaulmé, A.; Sarikaya, D.; Le Mut, K.; Despinoy, F.; Long, Y.; Dou, Q.; Chng, C.-B.; Lin, W.; Kondo, S.; Bravo-Sánchez, L.; Arbeláez, P.; Reiter, W.; Mitsuishi, M.; Harada, K.; and Jannin, P. 2021. MIcro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 212: 106452.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950. ArXiv: 1705.06950.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*,

volume 139 of *Proceedings of Machine Learning Research*, 5583–5594. PMLR.

Kim, Y.; Park, C.; Jeong, H.; Chan, Y. S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; and Park, H. W. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kong, Y.; and Fu, Y. 2022. Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, 130(5): 1366–1401.

Kovaleva, O.; Shivade, C.; Kashyap, S.; Kanjaria, K.; Wu, J.; Ballah, D.; Coy, A.; Karargyris, A.; Guo, Y.; Beymer, D. B.; and others. 2020. Towards visual dialog for radiology. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, 60–69.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, 2556–2563. Barcelona, Spain: IEEE. ISBN 978-1-4577-1102-2 978-1-4577-1101-5 978-1-4577-1100-8.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2022. F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models. Version Number: 2.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 1–10. Publisher: Nature Publishing Group.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Version Number: 2.

Li, Y.; Liu, M.; and Rehg, J. M. 2020. In the Eye of the Beholder: Gaze and Actions in First Person Video. ArXiv:2006.00626 [cs].

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. ArXiv:2310.03744 [cs].

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. Version Number: 2.

Liu, X.; Chen, K.; Wu, T.; Weidman, D.; Lure, F.; and Li, J. 2018. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Translational Research*, 194: 56–67.

Lukežič, A.; Vojíř, T.; Čehovin, L.; Matas, J.; and Kristan, M. 2018. Discriminative Correlation Filter with Channel and Spatial Reliability. *International Journal of Computer Vision*, 126(7): 671–688. ArXiv:1611.08461 [cs].

Mackenzie, C. F.; Yang, S.; Garofalo, E.; Hu, P. F.; Watts, D.; Patel, R.; Puche, A.; Hagegeorge, G.; Shalin, V.; Pugh, K.; Granite, G.; Stansbury, L. G.; Shackelford, S.; and Tisherman, S. 2021. Enhanced Training Benefits of Video Recording Surgery With Automated Hand Motion Analysis. *World Journal of Surgery*, 45(4): 981–987.

Mirchi, N.; Bissonnette, V.; Yilmaz, R.; Ledwos, N.; Winkler-Schwartz, A.; and Del Maestro, R. F. 2020. The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE*, 15(2): e0229596.

Novaes, M. d. A.; and Basu, A. 2020. Chapter 14 - Disruptive technologies: Present and future. In Gogia, S., ed., *Fundamentals of Telemedicine and Telehealth*, 305–330. Academic Press. ISBN 978-0-12-814309-4.

Nwoye, C. I.; Yu, T.; Gonzalez, C.; Seeliger, B.; Mascagni, P.; Mutter, D.; Marescaux, J.; and Padoy, N. 2022. Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos. *Medical Image Analysis*, 78: 102433. ArXiv:2109.03223 [cs].

Patocka, C.; Pandya, A.; Brennan, E.; Lacroix, L.; Anderson, I.; Ganshorn, H.; and Hall, A. K. 2024. The Impact of Just-in-Time Simulation Training for Healthcare Professionals on Learning and Performance Outcomes: A Systematic Review. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 19(1S): S32–S40.

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

Rojas, E.; Couperus, K.; and Wachs, J. 2020. The AI-Medic: an artificial intelligent mentor for trauma surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9: 1–9.

Roy, D.; Rajendiran, R.; and Fernando, B. 2023. Interaction Visual Transformer for Egocentric Action Anticipation. ArXiv:2211.14154 [cs].

Schmidt, A.; Sharghi, A.; Haugerud, H.; Oh, D.; and Mohareri, O. 2021. Multi-view Surgical Video Action Detection via Mixed Global View Attention. In De Bruijne, M.; Cattin, P. C.; Cotin, S.; Padoy, N.; Speidel, S.; Zheng, Y.; and Essert, C., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12904, 626–635. Cham: Springer International Publishing. ISBN 978-3-030-87201-4 978-3-030-87202-1. Series Title: Lecture Notes in Computer Science.

Shackelford, S. A.; Del Junco, D. J.; Riesberg, J. C.; Powell, D.; Mazuchowski, E. L.; Kotwal, R. S.; Loos, P. E.; Montgomery, H. R.; Remley, M. A.; Gurney, J. M.; and Keenan, S. 2021. Case-control analysis of prehospital death and prolonged field care survival during recent US military combat operations. *Journal of Trauma and Acute Care Surgery*, 91(2S): S186–S193.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Actor and Observer: Joint Modeling of First and Third-Person Videos. ArXiv:1804.09627 [cs].

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in

The Wild. *ArXiv*, abs/1212.0402. Publisher: arXiv Version Number: 1.

Stewart, T.; and Bird, P. 2022. Health economic evaluation: cost-effective strategies in humanitarian and disaster relief medicine. *BMJ Military Health*, e001859.

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.

Tao, L.; Elhamifar, E.; Khudanpur, S.; Hager, G. D.; and Vidal, R. 2012. Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation. In Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J. M.; Mattern, F.; Mitchell, J. C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B.; Sudan, M.; Terzopoulos, D.; Tygar, D.; Vardi, M. Y.; Weikum, G.; Abolmaesumi, P.; Joskowicz, L.; Navab, N.; and Jannin, P., eds., *Information Processing in Computer-Assisted Interventions*, volume 7330, 167–177. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-30617-4 978-3-642-30618-1. Series Title: Lecture Notes in Computer Science.

Ulhaq, A.; Akhtar, N.; Pogrebna, G.; and Mian, A. 2022. Vision Transformers for Action Recognition: A Survey. ArXiv:2209.05700 [cs, eess].

Valderrama, N.; Ruiz Puentes, P.; Hernández, I.; Ayobi, N.; Verlyck, M.; Santander, J.; Caicedo, J.; Fernández, N.; and Arbeláez, P. 2022. Towards Holistic Surgical Scene Understanding. In Wang, L.; Dou, Q.; Fletcher, P. T.; Speidel, S.; and Li, S., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13437, 442–452. Cham: Springer Nature Switzerland. ISBN 978-3-031-16448-4 978-3-031-16449-1. Series Title: Lecture Notes in Computer Science.

Vannaprathip, N.; Haddawy, P.; Schultheis, H.; and Suebnukarn, S. 2025. SDMentor: A virtual reality-based intelligent tutoring system for surgical decision making in dentistry. *Artificial Intelligence in Medicine*, 162: 103092.

Wachs, J. P.; Kirkpatrick, A. W.; and Tisherman, S. A. 2021. Procedural Telementoring in Rural, Underdeveloped, and Austere Settings: Origins, Present Challenges, and Future Perspectives. *Annual Review of Biomedical Engineering*, 23(1): 115–139.

Xiang, W.; Li, C.; Zhou, Y.; Wang, B.; and Zhang, L. 2023. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10276–10285.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2023. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. ArXiv:2311.06242 [cs].

Xu, M.; Islam, M.; and Ren, H. 2022. Rethinking Surgical Captioning: End-to-End Window-Based MLP Transformer Using Patches. ArXiv:2207.00113 [cs].

Xu, S.; Zhou, Y.; Liu, Z.; Wu, Z.; Zhong, T.; Zhao, H.; Li, Y.; Jiang, H.; Pan, Y.; Chen, J.; and others. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*.

Yan, T.; Zeng, W.; Xiao, Y.; Tong, X.; Tan, B.; Fang, Z.; Cao, Z.; and Zhou, J. T. 2024. Crossglg: Llm guides one-shot skeleton-based 3d action recognition in a cross-level manner. In *European Conference on Computer Vision*, 113–131. Springer.

Yang, J.; Dong, X.; Liu, L.; Zhang, C.; Shen, J.; and Yu, D. 2022. Recurring the Transformer for Video Action Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14043–14053.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; Liu, C.; Liu, M.; Liu, Z.; Lu, Y.; Shi, Y.; Wang, L.; Wang, J.; Xiao, B.; Xiao, Z.; Yang, J.; Zeng, M.; Zhou, L.; and Zhang, P. 2021. Florence: A New Foundation Model for Computer Vision. ArXiv:2111.11432 [cs].

Zhao, H.; Torralba, A.; Torresani, L.; and Yan, Z. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. ArXiv:1712.09374 [cs].

Özsoy, E.; Mamur, A.; Tristram, F.; Pellegrini, C.; Wysocki, M.; Busam, B.; and Navab, N. 2025. EgoExOR: An Ego-Exo-Centric Operating Room Dataset for Surgical Activity Understanding. ArXiv:2505.24287 [cs].