SAFER-Activities: A Dataset for Smart Assessment of Fall Events and Routine Activities

Diwas Lamsal[1]*, Pramod Wickramatilake[2], Jednipat Moonrinta[2], Mongkol Ekpanyapong[2], and Matthew N. Dailey[2]

[1] KU Leuven, Leuven 3000, Belgium
[2] Asian Institute of Technology, Khlong Luang, Pathum Thani 12120, Thailand

Abstract. Smart healthcare monitoring systems require precise action recognition to ensure well-being and timely intervention in critical situations such as falls, particularly for mobility-challenged individuals. Existing datasets are often clip-based, lacking the frame-level detail needed to recognize actions online, as they unfold. To address this, we introduce SAFER-Activities, a dataset for fall detection and physical activity monitoring, with a dedicated subset for wheelchair use scenarios. It comprises over 66 hours of video data captured by multiple cameras, with 85,310 action instances and frame-level annotations for 30 action classes. We benchmark action recognition on SAFER-Activities with 2D and 3D skeleton models, RGB models with frozen backbones, and multimodal fusion strategies, and evaluate on in-lab, out-of-distribution, and cross-dataset test sets. Skeleton-based models generalize best under domain shift; fusing frozen RGB features with the skeleton stream improves in-domain recognition over the baseline CNN1D, most clearly on the wheelchair subset, but degrades out of distribution. Cross-dataset and qualitative evaluations confirm that models trained on SAFER-Activities transfer well to unseen environments and external fall data. To support research on robust fall detection and activity monitoring, we release the dataset and code at `https://safer-activities.github.io/`.

Keywords: Action Recognition · Fall Detection · Skeleton-based Action Recognition · Wheelchair · Human Pose Estimation · Dataset

1 Introduction

Smart healthcare monitoring systems should be designed to analyze activities of daily living (ADL), ensuring that people maintain an adequate level of physical activity [7]. They should also enable timely intervention in critical situations like falls, which are a leading cause of injury-related hospitalizations [48]. Most smart healthcare monitoring systems are based on wearable sensors or cameras [4]. Wearable sensors can be uncomfortable for some users, and cognitively impaired individuals often forget to wear them [19]. Camera-based human action recognition (HAR) systems are effective complementary or standalone solutions.

* Work done while at the Asian Institute of Technology.

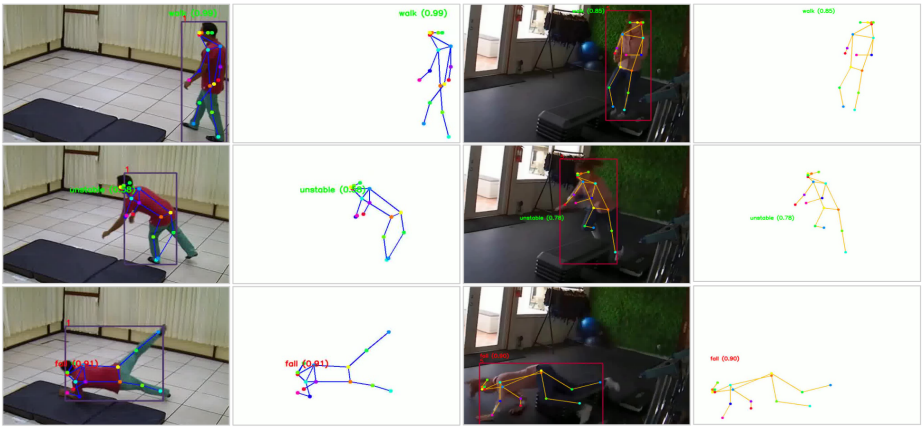

Fig. 1: Predictions from a 2D skeleton-based model trained on SAFER-Activities, evaluated on unseen in-lab test data (left) and a real-world fall example (right).

HAR is a central task in video understanding, with comprehensive datasets contributing to its progress in the last decade [22, 41]. Various features, including RGB [8, 45, 47], optical flow [8, 44], and human skeletons [16, 28, 50], have been explored for HAR. Skeleton-based methods offer concise representations of body movements that are robust to background clutter and lighting changes [16]. However, they discard scene context and can become unreliable when pose estimation is degraded by occlusion or unusual viewpoints [23]. RGB-based methods capture richer semantic information, including scene context and object interactions, but are more susceptible to domain shift, as appearance features learned from one environment often fail to transfer to new settings [46]. Combining both modalities is therefore an appealing direction [26].

HAR for smart healthcare monitoring requires datasets with long, untrimmed videos and precise frame-level annotations [27] rather than the short, coarsely labeled clips typical of large-scale activity recognition benchmarks [22,41]. These datasets must include a broad range of routine activities, such as walking and exercising, to enable physical activity monitoring, as well as actions resembling falls, such as lying down on a sofa, to reliably distinguish them from actual falls. They should also include evaluation data from unseen subjects and environments to verify that trained models generalize beyond the controlled training domain. While prior fall detection datasets [4, 9] feature realistic falls designed to mimic real-world scenarios, they are limited by a small number of fall instances and lack comprehensive labeling of routine and fall-like events.

Another gap in the activity recognition resources currently available lies in the unavailability of datasets containing wheelchair users. This group is particularly vulnerable and would greatly benefit from physical activity monitoring. Prolonged wheelchair use, especially propelling and lifting, can cause shoulder damage [13]. Wheelchair users also face heightened risks of falls during transfers [42]. Insufficient physical activity by some wheelchair users puts them at risk

of chronic diseases [5], hypertension, hyperlipidemia, and diabetes [38]. Therefore, developing robust activity recognition systems for this population is crucial for promoting their health, safety, and overall well-being.

To address these issues, we introduce SAFER-Activities: a large-scale dataset for fall detection and activity monitoring, with over 66 hours of video data from 46 participants, with frame-level annotations for 85,310 action instances across 30 classes, including 5,406 fall instances alongside routine and closely related actions, such as lying down and sitting. The dataset includes a subset focused on wheelchair use and a separate non-lab test set recorded in a home environment to evaluate real-world generalization. We also provide a complementary pose estimation dataset of 6,846 images and 8,324 human instances to benchmark pose estimation methods for people seated in wheelchairs.

We benchmark SAFER-Activities with a set of methods spanning 2D and 3D skeleton-based models, RGB-based models using pretrained frozen visual backbones, and multimodal fusion. Our evaluation covers in-lab, non-lab, and wheelchair test sets, as well as cross-dataset generalization on an external fall detection dataset. We find that while skeleton-based methods offer the strongest out-of-distribution generalization, multimodal fusion improves in-distribution performance but introduces challenges under domain shift, highlighting SAFER-Activities as a valuable testbed for robust action recognition and fall detection research. Models trained on SAFER-Activities effectively detect falls in unseen real-world scenarios (Fig. 1). Our contributions are as follows:

- We introduce SAFER-Activities, a large-scale dataset with dense frame-level annotations for falls and routine activities, multi-camera viewpoints, and a dedicated subset with wheelchair use scenarios, serving as a comprehensive benchmark for action recognition and fall detection.
- We provide extensive benchmarks spanning 2D pose, 3D pose via monocular lifting, frozen RGB-based pretrained backbones, and multimodal fusion methods across in-lab, out-of-distribution, and wheelchair evaluation splits.
- We include a non-lab test set recorded in a home environment with unseen participants and viewpoints, and perform cross-dataset evaluation on an external fall detection dataset.
- We release a complementary dataset to benchmark human pose estimation methods during wheelchair use.

2 Related Work

2.1 Fall Detection Datasets

Several small-scale datasets exist for vision-based fall detection (Tab. 1). Guerrero *et al.* [20] collect 50 fall instances from 10 participants in an uncontrolled environment with varying lighting, but label only fall versus non-fall. The ImViA dataset [9] features 222 videos with 143 fall instances across four realistic indoor settings; however, only fall boundaries are annotated. Martínez *et al.* [32] present UP-Fall, a multimodal dataset combining cameras with wearable and ambient

Table 1: Comparison of vision-based fall detection datasets. # Subj.: number of subjects. Age: age range of subjects. # Classes: total number of labeled activity categories (2 indicates fall/non-fall labels only). # Falls: total fall instances across all camera views. # ADL Inst.: total labeled non-fall activity instances (✗ if ADL are not individually labeled). # WC Falls: wheelchair fall instances (✗ if not included). # Hours: total duration of video data. N/A: not available. Only Auvinet *et al.* [3] and SAFER-Activities provide frame-level annotations for all activity classes.

Source	# Subj.	Age	# Classes	# Falls	# ADL Inst.	# WC Falls	# Hours
Guerrero *et al.* [20]	10	23–40	2	50	✗	✗	<1
Charfi *et al.* [9]	9	N/A	2	143	✗	✗	<1
Martínez *et al.* [32]	17	18–24	11	255	306	✗	∼ 9.4
Auvinet *et al.* [3]	1	N/A	9	200	1120	✗	∼ 10.6
Baldewijns *et al.* [4]	10	N/A	2	275	✗	20	∼ 41.4
SAFER-Activities	**46**	**18–58**	**30**	**5,406**	**79,904**	**840**	**>66**

sensors, providing clip-level annotations for 11 activity classes across 17 participants, all young adults aged 18–24. Auvinet *et al.* [3] provide 24 scenarios captured by eight cameras with frame-level labels for 9 activity categories and 1,120 ADL instances, but feature only a single participant. Baldewijns *et al.* [4] re-enact actual nursing-home falls across 72 scenarios from five camera angles, including 20 wheelchair fall instances, and emphasize realism with longer videos containing post-fall activities, but do not provide ADL labels.

Relative to prior vision-based fall-detection datasets, SAFER-Activities provides substantially more fall instances (5,406), frame-level annotations for 30 activity classes, and recordings from 46 adult participants spanning a broader age range (18–58). It also includes diverse routine and fall-like activities that cover typical scenarios encountered in realistic environments. Among the compared datasets, the ImViA dataset [9] is particularly suited for evaluating cross-dataset generalization, as it features a realistic home environment with variable lighting and precise fall boundary annotations. We use it for cross-dataset evaluation (see Tab. 6).

2.2 Camera-Based Human Action Recognition

Camera-based HAR commonly relies on RGB, skeleton, or multimodal representations. RGB methods capture appearance, scene context, and object interactions, ranging from temporal segment networks [47] and two-stream architectures [44] to 3D CNNs [18] and video transformers [2]. Recent pretrained models such as VideoMAE [45], CLIP [39], and DINOv3 [43] provide strong visual representations for image and video understanding. Skeleton-based methods instead operate on body joint coordinates, making them less sensitive to background and appearance changes, though dependent on pose quality. Graph-based models such as ST-GCN [50], MS-G3D [28], and DG-STGCN [14], as well as heatmap-based models such as PoseC3D [16], have shown strong perfor-

mance. 2D-to-3D lifting methods such as MotionAGFormer [33] further enable 3D skeleton recognition from monocular video.

Multimodal fusion methods combine complementary cues from RGB and skeleton streams. Simple approaches include feature concatenation and late score fusion [53], while more advanced methods balance modality learning through gradient modulation [37], input-quality weighting [51], modality dropout [34], or multi-modality co-learning [26]. However, robustness under domain shift remains a key challenge for both RGB-based models and fusion methods.

We evaluate these approaches on SAFER-Activities, which is comparable in scale to established HAR benchmarks such as NTU RGB+D [41] (56,880 instances, 40 subjects), PKU-MMD [25] (21,545 instances, 66 subjects), and Epic-Kitchens-100 [12] (89,977 instances, 37 subjects). It additionally features frame-level annotations, untrimmed multi-camera recordings, and dedicated in-distribution and out-of-distribution evaluation splits. Accordingly, our benchmark spans models ranging from a lightweight 1D CNN baseline to state-of-the-art skeleton architectures, paired with widely adopted pretrained RGB backbones used as frozen feature extractors.

2.3 Human Pose Estimation

The primary benchmarks for human pose estimation (HPE) include the MS COCO Keypoint Detection [24] and MPII Human Pose [1] datasets, which contain about 200,000 images (COCO) and 25,000 images (MPII) for pose estimation. The OCHuman dataset [52] aims to tackle the challenge of occlusion, with a collection of 5,081 images featuring heavily occluded humans. Likewise, the CrowdPose dataset [23] comprises about 20,000 images of humans in highly crowded scenes. SAFER-Activities features a specialized HPE data subset targeting people in wheelchairs, designed to benchmark HPE models' effectiveness in estimating their poses from diverse camera angles under occlusion.

3 Dataset

SAFER-Activities contains a collection of videos with frame-level annotations, tailored for action recognition and smart monitoring of people. It features a separate subset focusing on people in wheelchairs for action recognition and pose estimation. We describe the data collection methods, annotation procedures, and key statistics in this section; supplementary material provides additional details on the annotation tool, full action class definitions and illustrations, the visual feature extraction pipeline, and wheelchair pose estimation benchmarks.

3.1 Frame-level Labels

We define frame-level labels as those that precisely mark the start and end of an action amidst a sequence of actions. Frame-level labels allow models to exploit temporal dependencies between actions. Take, for instance, a classification

(a) Camera setup

(b) Example fall sequence

(c) Sample views and actions from in-lab non-wheelchair (left), wheelchair (center), and non-lab (right) subsets

Fig. 2: Room setup and sample actions

model M_1 tasked with predicting an action $A_t = M_1(F_{t-i}, \ldots, F_t, \ldots, F_{t+k})$ at a given time t, within a window spanning frames $F_{t-i}, \ldots, F_t, \ldots, F_{t+k}$, $(i, k \geq 0)$. Through such labeling and the use of untrimmed videos, not all frames in a given window are required to represent the same action, enabling the model to capture information from adjacent actions. Frame-level labels make it more feasible to train models for online action recognition, predicting current actions as they unfold using the temporal context provided by surrounding frames.

3.2 Dataset Construction

We performed data collection in four stages, each designed to capture a range of actions, from basic actions such as sitting, pointing, and clapping to critical actions such as falls. The second stage focused on the wheelchair subset, while the first, third, and fourth stages covered activities without wheelchairs. The setup for data collection for the first three stages featured a room with eight fixed high-resolution CCTV cameras aimed toward the center from different directions. We refer to these as the in-lab subsets. Stage 1 data were recorded at 25 fps with a resolution of 2688×1520, whereas for Stages 2, 3, and 4, we increased the resolution to 3072 × 2048. The fourth set (non-lab subset) was exclusively collected in an external home environment with a similar setup but using six cameras instead of eight. Upon entry to the room, participants performed a sequence of predefined activities, guided by audio instructions from a playback device. Figure 2a illustrates the setup of the cameras and room arrangement. Figure 2 provides some example actions and views taken from the dataset.

Ethics and Privacy. This study was approved by the research ethics review committee at the Asian Institute of Technology (Ref. No.: RERC 2022/011).

Participants reviewed and signed an informed consent form for their images to be publicly shared for research purposes prior to the data collection. They are free to ask for the removal of their personally identifiable information (face).

Data Annotation. To ensure precise action recognition, we developed an annotation tool to mark the start and end of each action in a video. The annotator was provided with a set of predefined labels to cover the expected actions and instructed to decide when each action starts and ends by watching the video closely, going back and forth if necessary to be certain. For example, a "fall" usually starts after a "stand" or "unstable" action when someone begins to fall, but the annotator must check the subsequent few seconds of the video to ensure that the action is a true fall. Once the person is on the ground, the "fall" action ends, and the "lie down" action begins. Following standard practice in temporally dense activity annotation [25], each video was labeled in its entirety by a single annotator to ensure temporal consistency across action boundaries, but independently verified by a second annotator who reviewed the labels and flagged any inconsistencies. Both annotators are computer science M.S. graduates working as researchers. To quantify reliability, this second annotator additionally re-annotated approximately 5% of the recordings from scratch, covering all action classes. Frame-level inter-annotator agreement reached a Cohen's κ of 0.89 (0.87 fall/non-fall, 0.87 non-wheelchair, 0.91 wheelchair), with a boundary F1 of 0.86 at a 500 ms tolerance and a median boundary deviation of 145 ms.

Actions. The final lists of actions we used for model training and analysis are provided in Fig. 4. In our experiments, we generalized common actions to focus the model on more precise fall detection and actions related to physical activity monitoring. For instance, various sitting actions such as clapping, checking the time, making phone calls, waving, and pointing are merged into a single "sit activity" macro action. This results in 30 unique actions across both non-wheelchair and wheelchair subsets, down to 15 each from the original 25 and 37, respectively. Descriptions for all the original actions, as well as the illustrations for the generalized macro actions are included in the supplement.

Pose Skeletons and Bounding Boxes. For skeleton-based action recognition, we extract the 2D poses used for training models, following the top-down approach with YOLOv8x [21] as the detector and the ViTPose-h-multi variant of ViTPose [49] as the pose estimator. The pose data are stored following the MMAction2 format [11]. We additionally provide 3D pose sequences obtained by lifting the detected 2D keypoints using a pretrained MotionAGFormer [33] network.

Visual Features. To complement the skeleton data, we extract visual features from person-centric crops using three frozen backbones: VideoMAE [45] for spatiotemporal features and CLIP [39] and DINOv3 [43] for frame-level features. Crops are obtained by tracking person bounding boxes across frames and resizing to 640×480. Features are extracted per-video and temporally aligned with the skeleton sequences to enable direct comparison and multimodal fusion.

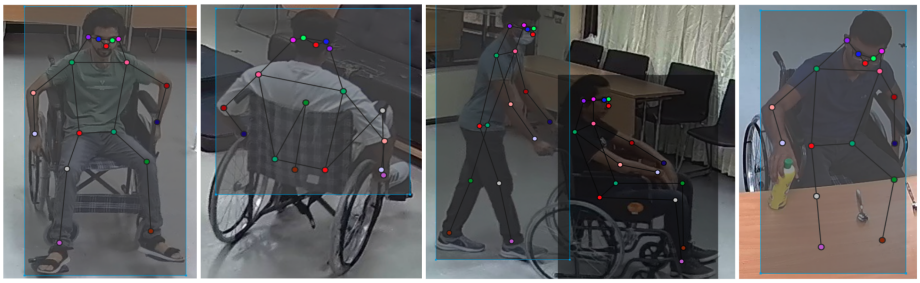

Fig. 3: Sample images from the wheelchair keypoints dataset. It contains annotations of people in wheelchairs from various angles and includes highly occluded scenes.

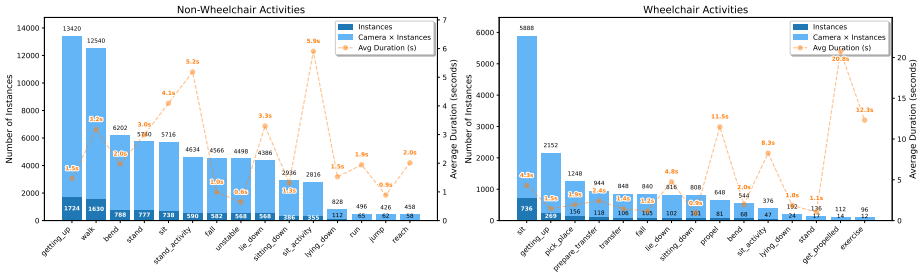

Fig. 4: Class statistics for the non-wheelchair (left) and wheelchair (right) subsets.

Wheelchair Keypoint Estimation. SAFER-Activities includes a complementary pose estimation dataset derived from the wheelchair video data (Fig. 3). It features 6,846 images containing 8,324 manually-annotated human instances. We used the COCO Annotator tool [6] to label person bounding boxes and keypoints. While we have not used this data to fine-tune any models, we do provide the performance of popular HPE methods on this subset. Future work could exploit these annotations to further improve the best keypoint detection models.

3.3 Statistics

SAFER-Activities comprises over 66 hours of labeled video data, with about 15 hours from the wheelchair subset. The data feature a diverse group of 46 participants. This includes 29 male and 17 female individuals, with an average age of approximately 31 years, ranging from 18 to 58. Figure 4 shows the total number of action instances and average action duration. "Camera × Instances" refers to the aggregation of multiple views, resulting in multiples of action instances.

Dataset Split Following Shahroudy *et al.* [41], we split both datasets by subject and camera view. The non-wheelchair dataset (467 videos) was divided into training/testing sets of 371/96 (subject-wise) and 350/117 (view-wise). The wheelchair dataset (142 videos) was split into training/testing sets of 96/46

(subject-wise) and 104/38 (view-wise). Additionally, 30 non-lab test videos from 5 participants were recorded in a home environment with different lighting conditions, camera viewpoints, and scenes with partial occlusion, none of which appear in the training data. This split serves as an out-of-distribution (OOD) evaluation to assess generalization beyond the controlled lab setting.

4 Experiments

We evaluate action recognition on SAFER-Activities using 2D and 3D skeleton models, frozen RGB-only models, and fusion approaches. Results are reported on the in-lab non-wheelchair, non-lab (OOD), and wheelchair test sets; subsequent tables label these splits as *In-Lab*, *Non-Lab*, and *Wheelchair*, respectively. We additionally perform cross-dataset evaluation on an external dataset and also analyze the effect of fusion strategies and the temporal context window size.

4.1 Experimental Setup

All models operate on the skeleton and visual features described in Sec. 3.2. We use the subject-wise splits and report macro-averaged and per-class F1-scores. We additionally report segment-level F1 at IoU thresholds in the supplement.

Input and label assignment. Following Duan *et al.* [16], a sliding window extracts 48-frame sequences from each video. During training, a sub-clip is randomly sampled every 20 frames; at test time, a dense sliding window covers the full video. Labels are assigned by majority vote over five frames centered on the window midpoint. The visual branch of frozen RGB and fusion models uses 16 frames, uniformly subsampled from the same 48-frame window.

Training protocol. For skeleton-based recognition, we evaluate a lightweight 1D CNN (CNN1D) on normalized joint coordinates, ST-GCN++ [15] and MS-G3D [28] on both 2D and lifted 3D poses, PoseC3D [16] on 2D, and DG-STGCN [14] on 3D. For RGB-only models, each frozen visual backbone is paired with a linear classifier; CLIP [39] and DINOv3 [43] features are temporally mean-pooled across frames, similar to Oquab *et al.* [36], while VideoMAE [45] directly produces a single clip-level representation. Multimodal fusion combines each visual backbone with the CNN1D skeleton stream via feature concatenation, keeping the skeleton branch fixed for a controlled comparison across fusion variants. The PYSKL framework [15] trains all GCN-based and PoseC3D skeleton models with SGD and cosine annealing [30], following the default PYSKL configurations, while RGB-only and fusion models use AdamW [31] with cosine annealing. Additional details for each model can be found in the training configuration files.

4.2 Baseline Results

Tables 2 and 3 report per-class F1-scores across all modalities and test sets. On the in-lab splits, 2D skeleton models achieve the highest overall scores, with MS-G3D reaching 87.2% on the non-wheelchair subset and 77.5% on the wheelchair

Table 2: Per-class F1 scores (%) on the non-wheelchair test sets. Top: in-lab, bottom: non-lab (OOD). **Bold** = best per column within each subset; underline = second best.

	Model	GU	W	B	S	SA	U	ST	F	LD	SIA	SD	LDN	R	RN	J	Avg
In-Lab Test Set																	
2D	CNN1D	88.7	94.1	84.5	83.8	96.7	71.8	87.5	92.3	93.4	91.3	79.5	56.9	70.3	70.9	94.9	83.8
2D	ST-GCN++	89.4	95.0	85.6	87.7	97.3	70.8	86.5	92.9	94.1	93.7	**85.0**	62.5	79.8	87.2	96.9	86.9
2D	PoseC3D	88.8	95.5	86.0	86.5	**97.9**	71.7	88.4	**93.6**	93.9	93.4	84.0	62.4	78.1	80.1	**98.3**	86.6
2D	MS-G3D	89.1	95.2	86.8	**89.0**	97.8	74.1	87.6	92.7	94.1	**95.1**	84.7	61.2	78.6	**88.9**	93.4	**87.2**
3D	ST-GCN++	88.8	94.0	84.3	83.0	97.0	48.6	87.3	91.8	94.1	90.6	80.6	68.9	67.7	80.4	82.9	82.7
3D	DG-STGCN	89.0	92.9	83.4	82.5	96.8	**74.8**	86.3	92.7	88.6	89.9	80.5	66.7	43.1	80.5	82.9	82.1
3D	MS-G3D	**89.8**	94.9	87.3	83.1	97.4	74.5	88.7	92.4	93.8	91.5	83.4	**69.3**	72.9	81.4	91.1	86.1
RGB	CLIP	65.3	91.5	79.2	79.7	94.6	33.9	79.6	84.9	91.1	93.5	40.2	16.6	65.2	63.9	80.3	70.6
RGB	DINOv3	69.9	92.9	81.5	78.3	94.0	35.9	78.9	87.8	91.8	90.2	47.6	19.6	72.1	61.8	79.7	72.1
RGB	VideoMAE	80.2	94.9	81.2	79.1	95.5	61.9	84.2	89.5	91.6	92.6	62.4	51.2	78.1	69.3	93.6	80.3
Fuse	CLIP	86.4	94.2	87.1	86.6	97.3	71.6	89.0	90.9	93.9	94.8	79.3	57.5	83.9	66.4	96.3	85.0
Fuse	DINOv3	87.8	**95.9**	**88.4**	85.7	97.7	73.0	**90.8**	92.3	**94.7**	93.4	82.4	59.0	81.9	65.7	95.0	85.6
Fuse	VideoMAE	88.6	95.0	88.1	87.9	97.4	70.9	89.7	92.4	94.1	94.7	81.4	56.0	**85.9**	69.7	95.1	85.8
Non-Lab Test Set (OOD)																	
2D	CNN1D	89.8	91.5	72.6	70.7	78.6	41.1	72.9	71.2	92.9	73.0	83.8	65.7	31.7	83.2	88.4	73.8
2D	ST-GCN++	92.1	93.0	75.7	**75.6**	81.6	43.1	73.4	**80.5**	**94.7**	74.6	87.5	67.3	26.6	90.6	94.8	76.7
2D	PoseC3D	91.1	**93.8**	**76.8**	75.0	**84.6**	**45.5**	**77.5**	79.3	94.6	**85.6**	87.6	60.5	48.1	**92.9**	**96.5**	**79.3**
2D	MS-G3D	90.9	93.7	72.2	73.7	83.9	41.1	75.1	80.1	93.2	77.9	**88.5**	70.1	37.7	91.7	95.8	77.7
3D	ST-GCN++	90.2	91.6	71.7	70.7	71.2	38.7	65.4	74.3	91.9	77.7	80.6	62.4	16.5	86.5	84.5	71.6
3D	DG-STGCN	90.7	90.0	67.6	71.3	74.3	38.2	67.6	78.1	92.9	75.3	81.2	**70.2**	32.0	79.2	79.9	72.6
3D	MS-G3D	**92.2**	92.2	73.2	70.9	75.5	44.4	69.5	80.3	92.5	72.7	85.4	60.5	**51.9**	85.3	83.8	75.3
RGB	CLIP	46.9	72.7	29.3	53.1	69.3	0.0	39.6	9.9	84.3	69.7	17.8	11.9	0.0	28.7	0.0	35.6
RGB	DINOv3	24.0	24.5	27.0	46.6	22.4	0.0	42.1	0.0	81.7	68.5	7.6	11.0	0.0	0.2	1.2	23.8
RGB	VideoMAE	69.7	79.2	26.0	55.9	68.8	37.4	56.0	49.0	84.7	69.3	39.5	17.1	22.8	25.4	30.4	48.7
Fuse	CLIP	81.7	81.7	51.7	58.9	76.5	15.5	59.2	58.8	89.9	72.6	59.8	54.7	3.3	19.8	0.0	52.3
Fuse	DINOv3	70.4	67.5	43.3	50.3	11.8	13.3	59.9	47.5	87.1	3.2	14.9	23.9	0.0	0.0	0.0	32.9
Fuse	VideoMAE	80.0	82.5	50.3	56.4	70.6	25.1	54.9	66.2	87.1	66.8	59.9	35.9	0.5	57.9	23.0	54.5

2D/3D = skeleton input (3D via MotionAGFormer [33] lifting), RGB = frozen pretrained features, Fuse = visual backbone + CNN1D feature concatenation; GU = Getting Up, W = Walk, B = Bend, S = Sit, SA = Standing Activity, U = Unstable, ST = Stand, F = Fall, LD = Lie Down, SIA = Sitting Activity, SD = Sitting Down, LDN = Lying Down, R = Reach, RN = Run, J = Jump.

subset. 3D pose models are competitive but slightly behind their 2D counterparts. The frozen RGB-only models lag substantially: VideoMAE is the strongest at 80.3% and 71.4%, while CLIP and DINOv3 trail further. Fusion is competitive with skeleton models in-lab (up to 85.8% non-wheelchair) and achieves the best wheelchair results, with DINOv3 and VideoMAE fusion both reaching 78.6%. Notably, all modalities detect falls reliably in controlled settings.

The non-lab results reveal a stark modality gap. Frozen RGB features collapse under domain shift: DINOv3 drops from 72.1% to 23.8%, CLIP from 70.6% to 35.6%, and VideoMAE from 80.3% to 48.7%. Skeleton models prove far more robust, with the best 2D score declining from 87.2% to 79.3%. Feature-concatenation fusion inherits the RGB weakness and falls below skeleton-only

Table 3: Per-class F1 scores (%) on the wheelchair test set. **Bold** = best per column; underline = second best.

	Model	S	GU	PP	PT	TR	F	LD	SD	PR	B	SIA	LDN	ST	GP	E	Avg
2D	CNN1D	86.1	77.3	56.5	65.6	55.2	76.7	89.1	70.9	86.2	67.5	70.3	41.9	29.1	67.0	83.1	68.2
	ST-GCN++	89.8	80.3	66.9	64.1	61.5	81.2	91.2	71.2	93.7	69.3	71.3	49.7	16.1	87.9	84.9	71.9
	PoseC3D	90.3	81.2	69.9	70.2	62.4	**87.0**	90.1	71.3	94.6	74.8	81.3	61.2	20.4	90.1	92.8	75.8
	MS-G3D	90.3	**82.5**	77.9	69.6	63.4	86.7	91.7	**74.9**	94.4	74.6	80.0	**67.1**	29.8	88.4	90.7	77.5
3D	ST-GCN++	87.7	80.1	70.1	68.8	63.9	81.8	92.5	70.5	87.1	73.7	51.3	55.3	22.8	92.3	87.6	72.4
	DG-STGCN	87.9	80.2	68.9	69.6	62.8	80.5	92.1	71.8	88.9	77.5	70.1	57.6	31.8	91.9	84.8	74.4
	MS-G3D	87.8	80.1	68.6	68.3	63.9	81.9	73.4	74.8	87.8	74.7	71.8	64.0	32.3	92.1	88.3	74.0
RGB	CLIP	86.9	48.9	65.3	41.2	30.1	41.3	89.9	44.4	86.8	62.4	66.9	29.7	25.4	95.6	88.2	60.2
	DINOv3	87.7	51.5	58.2	53.6	55.8	59.5	91.0	44.3	90.1	64.0	61.4	28.6	36.5	95.9	91.1	64.6
	VideoMAE	89.6	68.4	77.6	55.7	60.6	78.0	90.8	53.7	**96.1**	74.4	78.5	35.4	21.8	**99.4**	91.4	71.4
Fuse	CLIP	89.4	79.6	69.1	63.9	62.2	80.3	94.1	72.3	88.9	73.2	76.7	55.2	37.3	93.9	90.4	75.1
	DINOv3	91.4	80.0	72.5	**72.4**	**67.7**	80.4	94.1	69.6	92.6	82.7	76.0	58.4	**50.7**	94.5	**95.4**	**78.6**
	VideoMAE	**92.2**	82.1	**80.5**	69.5	66.6	80.3	**94.9**	74.0	95.9	**83.9**	**81.4**	50.3	38.5	99.3	90.4	**78.6**

2D/3D = skeleton input (3D via MotionAGFormer [33] lifting), RGB = frozen pretrained features, Fuse = visual backbone + CNN1D feature concatenation; S = Sit, GU = Getting Up, PP = Pick/Place, PT = Prepare Transfer, TR = Transfer, F = Fall, LD = Lie Down, SD = Sitting Down, PR = Propel, B = Bend, SIA = Sitting Activity, LDN = Lying Down, ST = Stand, GP = Get Propelled, E = Exercise.

performance. The disparity is most severe for fall detection: RGB models nearly fail entirely (DINOv3 0.0% F1, CLIP 9.9%), whereas skeleton models maintain 70–80% F1.

Several classes remain challenging across all models. "Unstable" is consistently the hardest (best in-lab: 74.8%, best OOD: 45.5%), likely due to its short average duration and motion patterns that overlap with other actions. "Lying Down" and "Reach" also degrade sharply under domain shift, the latter possibly confused with "Bend" across viewpoints. The gap between the lightweight CNN1D and state-of-the-art models widens in these harder settings, indicating that more expressive architectures better capture the fine-grained temporal and spatial cues needed for robust recognition. These findings point to three open challenges for future work: improving OOD robustness, fusion strategies that do not inherit the visual domain gap, and disambiguating visually similar actions.

4.3 Additional Experiments

Fusion strategies. Section 4.2 showed that feature concatenation inherits the frozen RGB domain-shift weakness. Table 4 compares alternative fusion strategies introduced to mitigate issues in modality fusion across all three visual backbones. ModDrop [34] is the most effective, substantially recovering non-lab performance: VideoMAE improves from 54.5% to 67.3%, DINOv3 from 32.9% to 59.6%, and CLIP from 52.3% to 61.9%. QMF [51] and OGM-GE [37] do not consistently improve over concatenation. MMCL [26], which discards RGB at inference and uses only skeleton, does not improve non-lab performance over the

Table 4: Macro F1 (%) for fusion strategies pairing each visual backbone with CNN1D skeleton features. Dropout = ModDrop [34]; MMCL [26] uses skeleton only at inference. **Bold** = best, underline = second best.

	Backbone	In-Lab	Non-Lab	Wheelchair
	CNN1D (skel. only)	83.8	**73.8**	68.2
Concat	DINOv3	85.6	32.9	**78.6**
	CLIP	85.0	52.3	75.1
	VideoMAE	85.8	54.5	**78.6**
Dropout	DINOv3	85.4	59.6	75.4
	CLIP	85.4	61.9	74.3
	VideoMAE	**86.3**	67.3	76.3
QMF	DINOv3	85.6	32.5	76.6
	CLIP	85.5	52.6	75.2
	VideoMAE	86.2	52.4	78.1
OGMGE	DINOv3	85.5	29.0	78.4
	CLIP	83.4	45.3	75.9
	VideoMAE	86.0	55.2	78.5
MMCL	DINOv3	82.5	66.6	68.9
	CLIP	82.2	68.0	69.7
	VideoMAE	83.4	66.5	68.7

Table 5: Temporal stride (TS) effect on 2D skeleton models (macro F1, %). **Bold** = best per column.

	In-Lab		Wheelchair	
Model	TS 1	TS 3	TS 1	TS 3
CNN1D	83.8	84.3	68.2	69.3
ST-GCN++	86.9	85.7	71.9	75.6
PoseC3D	86.6	86.7	75.8	78.3
MS-G3D	**87.2**	**87.1**	**77.5**	**78.5**

Table 6: Cross-dataset fall detection on ImViA [9] without fine-tuning. **Bold** = best per column.

	Model	Precision	Recall	F1
2D	CNN1D	92.0	94.1	93.0
	PoseC3D	98.9	94.8	**96.8**
RGB	CLIP	undef.	0.0	0.0
	DINOv3	undef.	0.0	0.0
	VideoMAE	86.4	38.4	53.1
Fuse	CLIP	**100.0**	66.7	80.0
	DINOv3	**100.0**	64.6	78.5
	VideoMAE	96.9	**96.0**	96.4

CNN1D baseline but notably improves wheelchair recognition (e.g., 69.7% vs. 68.2% for CNN1D alone). Despite these strategies, no fusion method surpasses the skeleton-only CNN1D on the non-lab set (73.8%), indicating that better fusion strategies are needed to leverage the strengths of each modality under domain shift.

Temporal context window. Table 5 compares temporal stride 1 (48 consecutive frames) and stride 3 (sampling every 3rd frame, covering 144 frames) for the 2D skeleton models. In-lab non-wheelchair performance is nearly identical across strides, but wheelchair recognition improves consistently with a longer context.

Cross-dataset fall detection. To evaluate cross-dataset generalization, we test a subset of models trained on SAFER-Activities on the ImViA dataset [9], using the 130 videos (99 falls) that provide frame-level fall annotations. Since ImViA contains only fall annotations, we evaluate fall detection exclusively. Poses and RGB features are extracted using the same pipeline described in Sec. 3.2. To assess accuracy, we cluster consecutive fall predictions into temporal events and match them against ground-truth clusters; a tolerance window of 15 frames accounts for labeling differences across datasets. Table 6 shows that skeleton models generalize well: PoseC3D achieves 96.8% F1 and CNN1D reaches 93.0%. Frozen RGB-only models confirm the domain-shift vulnerability: CLIP and DINOv3 predict no fall events at all, so their precision is undefined and their recall and

F1 are 0; VideoMAE, by contrast, suffers from very low recall (38.4%). Fusion of VideoMAE with CNN1D recovers this gap, nearly matching PoseC3D at 96.4% F1, whereas CLIP and DINOv3 fusion achieve high precision but limited recall.

4.4 Qualitative Analysis

Figure 5 shows CNN1D predictions on diverse data, including external real-world falls and wheelchair user actions. On external data (rows 1–5), skeleton-based models trained on SAFER-Activities perform well outside the training distribution, with failures primarily due to poor-quality poses under severe occlusion or missed human detections. Notably, the frame-level temporal annotations in SAFER-Activities enable even a lightweight model such as CNN1D to distinguish falls from visually similar actions like slowly lying down on a surface (row 6). This distinction, raised as a key concern for fall detection systems, relies on the abrupt temporal dynamics rather than static pose similarity of actions.

5 Conclusion

We presented SAFER-Activities, a large-scale dataset with frame-level labels for online action recognition, especially fall detection, in untrimmed videos. The dataset includes a dedicated wheelchair subset and is released with precomputed pose skeletons and visual features, providing a comprehensive resource for smart healthcare monitoring research. Our evaluation spans 2D and 3D skeleton models, frozen RGB backbones, and multimodal fusion. Skeleton-based models prove the most robust under domain shift, generalizing well to an out-of-distribution home environment and an external fall dataset. Fusing frozen RGB features with the baseline CNN1D improves in-domain recognition, most clearly on the wheelchair subset, but degrades out of distribution; the tested fusion strategies reduce but do not fully close this gap. The dense temporal annotations enable even lightweight models to distinguish visually similar actions such as falls and lying down based on temporal dynamics. We believe that SAFER-Activities, with its multi-modality benchmarks and out-of-distribution evaluation, will stimulate research on robust action recognition for safety-critical healthcare applications.

Limitations and Future Work. For safety and ethics reasons, fall simulations were performed by adult actors rather than elderly participants or regular wheelchair users. This design enabled controlled, repeatable capture of diverse fall and fall-like motions, while prospective validation with the intended populations remains an important next step. SAFER-Activities is therefore best viewed as a pre-deployment benchmark for fall dynamics, fall-like routine activities, and wheelchair-use scenarios rather than a clinical validation study. The wheelchair subset focuses on in-lab recordings; without a public external wheelchair activity dataset, our current external evidence for wheelchair-use scenarios is qualitative (Fig. 5). Future expansions with elderly participants, naturalistic falls, and larger non-lab wheelchair-use recordings would further strengthen ecological validity.

Fig. 5: CNN1D predictions on external and in-lab data. Rows 1–2: real falls, row 1 right and row 2 from [17]. Row 3: realistic fall datasets. Row 4: failures from missed detections (left) and severe occlusion (right). Row 5: wheelchair actions from [40] (fall, left) and [29] (transfer, right). Row 6 (in-lab): fall (left) versus slowly lying down (right).

Our findings suggest several directions for model development. First, our synchronized recordings enable robust multi-view architectures [35]. Second, because cross-dataset evaluations reveal that frozen RGB features struggle with appearance shifts, future fusion strategies must adaptively leverage visual context when it is reliable, while defaulting to robust pose dynamics under domain shift. Finally, sharp performance drops on visually ambiguous, safety-critical actions like falls motivate methods that combine pose dynamics with complementary depth or inertial cues to disambiguate difficult cases [10].

Acknowledgement. This work was supported by Thailand's office of the National Broadcasting and Telecommunications Commission and the Broadcasting and Telecommunications Research and Development Fund for Public Interest under Grant A64-1-(2)-006.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3686–3693 (2014). https://doi.org/10.1109/CVPR.2014.471
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A video vision transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6816–6826 (2021). https://doi.org/10.1109/ICCV48922.2021.00676
3. Auvinet, E., Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Multiple cameras fall dataset. Technical Report 1350, DIRO-Université de Montréal (2010)
4. Baldewijns, G., Debard, G., Mertes, G., Vanrumste, B., Croonenborghs, T.: Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. Healthcare Technology Letters **3**(1), 6–11 (2016). https://doi.org/10.1049/htl.2015.0047
5. Booth, F.W., Roberts, C.K., Laye, M.J.: Lack of exercise is a major cause of chronic diseases. Comprehensive Physiology **2**(2), 1143–1211 (2012). https://doi.org/10.1002/j.2040-4603.2012.tb00425.x
6. Brooks, J.: COCO Annotator. https://github.com/jsbroks/coco-annotator/ (2019), last accessed 2024/06/15
7. Bull, F.C., Al-Ansari, S.S., Biddle, S., Borodulin, K., Buman, M.P., Cardon, G., Carty, C., Chaput, J.P., Chastin, S., Chou, R., Dempsey, P.C., DiPietro, L., Ekelund, U., Firth, J., Friedenreich, C.M., Garcia, L., Gichu, M., Jago, R., Katzmarzyk, P.T., Lambert, E., Leitzmann, M., Milton, K., Ortega, F.B., Ranasinghe, C., Stamatakis, E., Tiedemann, A., Troiano, R.P., van der Ploeg, H.P., Wari, V., Willumsen, J.F.: World Health Organization 2020 guidelines on physical activity and sedentary behaviour. British Journal of Sports Medicine **54**(24), 1451–1462 (2020). https://doi.org/10.1136/bjsports-2020-102955
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the Kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.502
9. Charfi, I., Miteran, J., Dubois, J., Atri, M., Tourki, R.: Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification. Journal of Electronic Imaging **22**(4), 041106 (2013). https://doi.org/10.1117/1.JEI.22.4.041106
10. Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 168–172 (2015). https://doi.org/10.1109/ICIP.2015.7350781
11. Contributors, M.: OpenMMLab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2 (2020), last accessed 2024/06/15
12. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. International Journal of Computer Vision (IJCV) **130**, 33–55 (2022). https://doi.org/10.1007/s11263-021-01531-2

13. Ding, D., Hiremath, S., Chung, Y., Cooper, R.: Detection of wheelchair user activities using wearable sensors. In: Stephanidis, C. (ed.) Universal Access in Human-Computer Interaction. Context Diversity. pp. 145–152. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
14. Duan, H., Wang, J., Chen, K., Lin, D.: DG-STGCN: Dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv preprint arXiv:2210.05895 (2022)
15. Duan, H., Wang, J., Chen, K., Lin, D.: PYSKL: Towards good practices for skeleton action recognition. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 7351–7354 (2022)
16. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.00298
17. FailArmy: Failarmy's youtube channel. YouTube (2025), https://www.youtube.com/@failarmy, last accessed 2025/05/15
18. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (Oct 2019). https://doi.org/10.1109/iccv.2019.00630
19. Fleming, J., Brayne, C.: Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. BMJ **337** (2008). https://doi.org/10.1136/bmj.a2227
20. Guerrero, J.C.E., España, E.M., Añasco, M.M., Lopera, J.E.P.: Dataset for human fall recognition in an uncontrolled environment. Data in Brief **45**, 108610 (2022). https://doi.org/10.1016/j.dib.2022.108610
21. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Jan 2023), https://github.com/ultralytics/ultralytics, last accessed 2024/06/15
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The Kinetics human action video dataset (2017), https://arxiv.org/abs/1705.06950
23. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2019). https://doi.org/10.1109/cvpr.2019.01112
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context, p. 740–755. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In: Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities. p. 1–8. VSCC '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3132734.3132739
26. Liu, J., Chen, C., Liu, M.: Multi-modality co-learning for efficient skeleton-based action recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 4909–4918. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3681015
27. Liu, Y., Wang, L., Wang, Y., Ma, X., Qiao, Y.: FineAction: A fine-grained video dataset for temporal action localization. IEEE Transactions on Image Processing **31**, 6937–6950 (2022). https://doi.org/10.1109/tip.2022.3217368

28. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 140–149. IEEE Computer Society, Los Alamitos, CA, USA (jun 2020). https://doi.org/10.1109/CVPR42600.2020.00022
29. LiveToRoll: Livetoroll's youtube channel. YouTube (2025), https://www.youtube.com/livetoroll, last accessed 2025/05/15
30. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
32. Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., Peñafort-Asturiano, C.: UP-Fall detection dataset: A multimodal approach. Sensors **19**(9) (2019). https://doi.org/10.3390/s19091988
33. Mehraban, S., Adeli, V., Taati, B.: MotionAGFormer: Enhancing 3D human pose estimation with a Transformer-GCNFormer network. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024). https://doi.org/10.1109/WACV57701.2024.00677
34. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: ModDrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1692–1706 (2016). https://doi.org/10.1109/TPAMI.2015.2461544
35. Nguyen, T.T., Kawanishi, Y., John, V., Komamizu, T., Ide, I.: MultiSensor-Home: A wide-area multi-modal multi-view dataset for action recognition and transformer-based sensor fusion. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10 (2025). https://doi.org/10.1109/FG61629.2025.11099071
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
37. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8228–8237 (2022). https://doi.org/10.1109/CVPR52688.2022.00806
38. Popp, W.L., Richner, L., Brogioli, M., Wilms, B., Spengler, C.M., Curt, A.E.P., Starkey, M.L., Gassert, R.: Estimation of energy expenditure in wheelchair-bound spinal cord injured individuals using inertial measurement units. Frontiers in Neurology **9** (Jul 2018). https://doi.org/10.3389/fneur.2018.00478
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
40. regor09: regor09's youtube channel. YouTube (2025), https://www.youtube.com/@regor09, last accessed 2025/05/15
41. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A large scale dataset for 3D human activity analysis. In: 2016 IEEE Conference on Computer Vision

and Pattern Recognition (CVPR). IEEE (Jun 2016). https://doi.org/10.1109/cvpr.2016.115
42. Sheikh, S.Y., Jilani, M.T.: A ubiquitous wheelchair fall detection system using low-cost embedded inertial sensors and unsupervised one-class SVM. Journal of Ambient Intelligence and Humanized Computing **14**(1), 147–162 (Apr 2021). https://doi.org/10.1007/s12652-021-03279-6
43. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S.E., Ramamonjisoa, M., Massa, F., HAZIZA, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jegou, H., Labatut, P., Bojanowski, P.: DINOv3. Transactions on Machine Learning Research (2026), featured Certification
44. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. p. 568–576. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
45. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
46. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528 (2011). https://doi.org/10.1109/CVPR.2011.5995347
47. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(11), 2740–2755 (2019). https://doi.org/10.1109/TPAMI.2018.2868668
48. World Health Organization: WHO Global Report on Falls Prevention in Older Age. World Health Organization (2008), https://iris.who.int/handle/10665/43811, last accessed 2024/11/15
49. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: simple vision transformer baselines for human pose estimation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
50. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018)
51. Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J.T., Peng, X.: Provable dynamic fusion for low-quality multimodal data. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23 (2023)
52. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2Seg: Detection free human instance segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 889–898 (2019). https://doi.org/10.1109/CVPR.2019.00098
53. Zhu, X., Zhu, Y., Wang, H., Wen, H., Yan, Y., Liu, P.: Skeleton sequence and RGB frame based multi-modality feature fusion network for action recognition. ACM Trans. Multimedia Comput. Commun. Appl. **18**(3) (Mar 2022). https://doi.org/10.1145/3491228