scientific reports

Check for updates

OPEN

Reducing annotation burden in physical activity research using vision language models

Abram Schönfeldt[1], Benjamin Maylor[1], Xiaofang Chen[2], Ronald Clark[3] & Aiden Doherty[1]✉

Data from wearable devices collected in free-living settings, and labelled with physical activity behaviours compatible with health research, are essential for both validating existing wearable-based measurement approaches and developing novel machine learning approaches. One common way of obtaining these labels relies on laborious human annotation of sequences of images captured by body-worn cameras. The aim of this study was to investigate whether open-source vision-language models could accurately annotate activity intensity classes in wearable camera-based validation studies, thereby reducing the annotation burden. We compared the performance of three vision language models and two discriminative models on two free-living validation studies with 161 and 111 participants, collected in Oxfordshire, United Kingdom and Sichuan, China, respectively, using the Autographer (OMG Life, defunct) wearable camera. We found that the best open-source vision-language model (VLM) and fine-tuned discriminative model (DM) achieved comparable performance when predicting sedentary behaviour from single images on unseen participants in the Oxfordshire study; median F_1-scores: VLM = 0.89 (0.84, 0.92), DM = 0.91 (0.86, 0.95). Performance declined for light [VLM = 0.60 (0.56, 0.67), DM = 0.70 (0.63, 0.79)], and moderate-to-vigorous intensity physical activity [VLM = 0.66 (0.53, 0.85); DM = 0.72 (0.58, 0.84)]. When applied to the external Sichuan study, performance fell across all intensity categories, with median Cohen's κ scores falling from 0.54 (0.49, 0.64) to 0.26 (0.15, 0.37) for the VLM, and from 0.67 (0.60, 0.74) to 0.19 (0.10, 0.30) for the DM. Freely available computer vision models could help annotate sedentary behaviour, typically the most prevalent activity of daily living, from wearable camera images within similar populations to seen data, reducing the annotation burden when using cameras as the source of ground-truth.

Wearable measurements of physical activity behaviours have helped advance our understanding of the relationship between physical activity and health outcomes[1], provided more sensitive outcomes in clinical trials[2] and introduced new ways of monitoring population physical activity levels[3]. The most realistic setting for validating behaviour measurement approaches and developing novel machine learning approaches[4–8] is in diverse populations of people living their everyday lives, highlighting the need for large, labelled, wearable data-sets, captured in free-living conditions[9–11].

Activity intensity classes, Sedentary Behaviour (SB), Light Intensity Physical Activity (LIPA) and Moderate-to-Vigorous Physical Activity (MVPA), provide a simple classification of daily activities based on their energy expenditure, are clearly defined[13–15], and have been widely adopted in epidemiological research[6,16,17] and physical activity guidelines[18]. A pragmatic approach to collecting these data-sets in free-living settings has been for participants to wear cameras, which record footage that later is reviewed by annotators to inform the ground-truth labels[11,19]. However, the sensitive nature of this footage has meant that access to it is restricted to select researchers, trained to handle sensitive data[20], making it costly and time-consuming to label.

Recently, Keadle et al.[15] proposed adopting approaches from computer vision to predict aspects of physical activity in a study of 26 adults, using video-recorded direct observation, emphasising the distinction between the definitions of physical activity used in health research[13,21,22], such as activity intensity, and the varied definitions of activity prevalent in human activity recognition literature[23]. This work estimates the performance of computer vision methods based on video-recorded direct observation, leaving the performance on studies using wearable image-capturing cameras unexplored, in addition to questions of how stable model performance will be between different populations, and within larger populations.

[1]Department of Population Health, University of Oxford, Oxford, UK. [2]School of Epidemiology and Health Statistics, Chengdu Medical College, Sichuan, China. [3]Department of Computer Science, University of Oxford, Oxford, UK. ✉email: aiden.doherty@ndph.ox.ac.uk

In this work, we evaluate using open-source Vision Language Models (VLMs), and Discriminative Models (DMs) to classify activity intensity in two validation studies collected in Oxfordshire, United Kingdom[19] and Sichuan, China[24], with wearable camera data from 161 and 111 participants respectively (Fig. 1). Although ethical issues prevent us from making the wearable camera portion of these data-set publicly available, a detailed quality assessment of these data-sets is conducted, and we will make our codebase and models publicly available (the annotated wrist-worn accelerometer data is publicly available for the Oxfordshire study[19]). To our knowledge, this is the first work which assesses the adoption of VLMs in this setting, and highlights a less labour-intensive approach to gathering labelled validation data-sets in free-living settings.

Relevant work

Measuring activity intensity at scale

In this paper, we focus on the activity intensity classes *sedentary behaviour*, *light intensity physical activity* and *moderate-vigorous physical activity*, which are defined as:

Sedentary behaviour (SB)	waking behaviour at $\leq$1.5 METs in a sitting, lying or reclining posture,
Light intensity physical activity(LIPA)	waking behaviour at $<$ 3METs not meeting the sedentary behaviour definition,
Moderate-vigorous physical activity(MVPA)	waking behaviour at $\geq$ 3METs, and
Sleep	Non-waking behaviour (not used in this work, though included for completeness),

where the metabolic equivalent of task (MET) index estimates the ratio of an activity's metabolic rate to a resting metabolic rate, set by convention to 3.5 ml O_2 kg body weight^{-1} min^{-1}[14,25]. These definitions of the activity intensity classes are in line with the definition of SB obtained through consensus in[13], and the definitions of LIPA and MVPA used by[14,15]. Current WHO guidelines on physical activity are framed in terms of these activity intensity classes, emphasising the importance of accurately monitoring activity intensity in large populations[18].

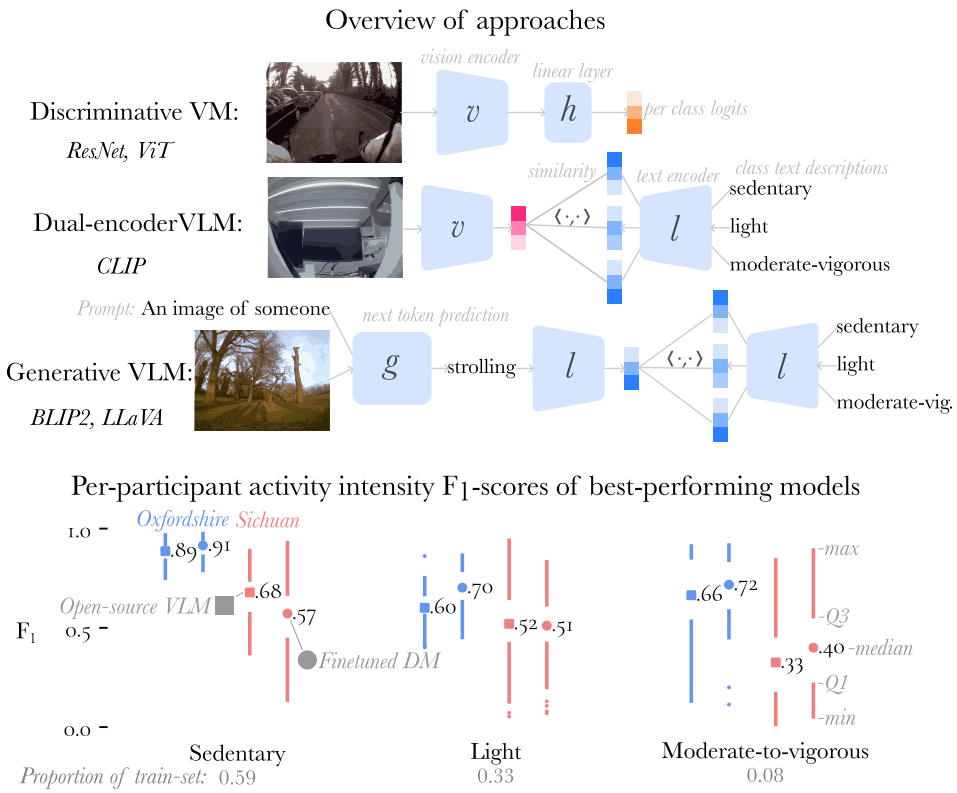


Fig. 1. Illustration of the computer vision approaches compared (top). Below, quartile plots[12] show the five-number summary of per-participant F_1-scores for sedentary behaviour (SB), light intensity physical activity (LIPA), and moderate-to-vigorous physical activity (MVPA), for the best-performing vision-language model, LLaVA (squares), and the best-performing discriminative vision model, ViT (circles), selected via hyperparameter tuning. Performance is shown for participants in the Oxfordshire study (blue) and the Sichuan study (red) withheld from model selection. MVPA constitutes only 8% of the training set, which is reflected in the high variance of per-participant F_1-scores.

Although wearable cameras alone could be used to measure activity intensity classes, they raise privacy concerns, making deploying them at scale less feasible. Thus, wearable cameras play a supporting role in capturing concurrent data that can be labelled for training activity intensity recognition models based on other wearable devices[5,26], and assessing the performance of these approaches in free-living settings, in so-called validation studies[19]. In[10]'s framework for developing devices that measure physical behaviour, they recognise both laboratory development (phase I), semi-structured evaluation (phase II) and naturalistic evaluation (phase III) as being necessary. Within this framework, wearable cameras can contribute naturalistic labelled data. This supports both the development of the new machine learning-based approaches, and the naturalistic evaluation of devices (phase III).

Although machine learning based approaches to measuring activity intensity classes from wearable accelerometers have overcome many of the shortcomings of than approaches reliant on cutpoints[27–29], they still have issues generalising to unseen populations[24], supporting the need to collect more training and validation data in new populations.

Currently, indirect calorimetry is recognised as the best way of measuring activity intensity over time[10]. In order to capture indirect calorimetry outside of lab-based conditions, participants wear facemasks and backpacks measuring the volume of oxygen inspired and expired over time. However, it is difficult for participants to wear these for extended periods of time in free-living situations. Doubly labelled water is a means of measuring total energy expenditure, but it is not possible to determine when participants were engaged in activities of different intensities using doubly labelled water[30].

Since indirect calorimetry is impractical in free-living settings over long durations, a pragmatic alternative is to use wearable camera footage to annotate activity intensity at scale[10]. In this approach, footage captured by the participant is reviewed by a trained annotator, and based on which activity the participant is engaged in, the annotator is able to estimate the corresponding activity intensity class using the Compendium of Physical Activity as reference. This approach has been used in a number of free-living validation studies[19,31–34], and has been shown to relatively accurate compared to indirect calorimetry[35].

Wearable data-sets of health-relevant behaviours

There are varying approaches to capturing free-living data-sets using cameras, arising from where the cameras are positioned relative to the participants, and the frequency with which cameras capture frames. Cameras can be worn by the participants, resulting in *egocentric* footage, held by observers following the participants, or placed in static positions, with the latter two options resulting in *third-person* footage. The frame-rate can be high, as is the case with video, or low, resulting in sparse sequences of images, similar to a time-lapse. Historically, battery limitations have meant that there has been a trade-off between the temporal resolution, and total duration recorded. For instance, Keadle et al.[15] used GoPros to record two sessions of 2 h of free-living data in a study of 26 participants. On the other hand, the studies considered in this work have recordings covering 8+ h in over 100 participants each, though at the expense of only capturing images every 20+ s. In Table 1, we highlight the sizes of comparable camera based validation studies, and there is a notable gap between the size of studies achieved using video compared to time-lapse recordings.

CAPTURE24: the Oxfordshire and Sichuan studies

The CAPTURE24 study was collected in 2014 from 165 participants in the Oxfordshire county of the United Kingdom in order to validate wrist-worn accelerometer-based physical activity measurement approaches in adults[19,42]. The CAPTURE24-CN study was collected in 2017 from 113 participants in the Sichuan province of China alongside a similar effort to develop and validate approaches to derive wrist-worn accelerometer-based physical activity measurements in over 20,000 participants in the China Kadoorie Biobank[24]. Though these studies only comprise roughly 100 participants each, they are the primary source of labelled data used to validate the measurements conducted in large scale health studies such as the UK and China Kadoorie Biobank[6–8,43],

Viewpoint	No. participants	Median δt (s)	Hours labelled	Paper
1st	161	24	1546	[19] (Oxfordshire)
1st	111	84	1078	[24] (Sichuan)
1st	50	15	1218	[33]
1st	22	Video	11	[37]
3rd	22	Video	34.3	[38]
1st	25	20	768	[34]
1st	22	Video	38	[39]
3rd	48	Video	192	[40]
3rd	31	Video	31	[41]

Table 1. Number of participants and estimated number of labelled hours of studies using cameras to validate wearable measurements of physical activity identified in a recent systematic review[36], and scoping review[35]. We recommend referring to the reviews for a more comprehensive list of validation studies. The two studies collected in Oxfordshire and Sichuan used in this work are shown at the top of this table. The estimates of the number of hours of labelled data for the timelapse studies is optimistic, since the temporal resolution of timelapse is much lower than video, resulting in periods of time that are difficult to label.

comprising tens of thousands of participants. As highlighted in Table 1, they represent the largest available validation studies.

Recognising activities from sparse sequences of egocentric images

Collecting and analysing data using wearable cameras has a history spanning over 3 decades, with pioneering work by Mann[44] and Aizawa[45], but was also foreseen as early as 1945[46]. There have been several works which explore human activity recognition in third person[15,47–49], and, to a lesser extent, egocentric videos[50–52]. Working towards the goal of reducing annotation burden in wearable data-sets, Bock et al.[53] proposed a clustering-based strategy where annotators label a representative clip in clusters of similar clips, derived from vision-foundation model features[54–56], which is then applied to all clips within each cluster. In contrast, we focus on methods which do not require human input, and which work on sparse sequences of images.

There has been some prior work on human activity recognition from sparse, egocentric sequences of images[57–62], though in datasets with only 10s of participants. These works focus on training discriminative models to predict predefined sets of labels, but the variation in how these labels are defined, and lack of publicly available benchmarks, makes it difficult to compare results across different works.

Though there has been less work on modelling activity from sparse sequences of egocentric images seems over the past few years, there has been increased interest in modelling egocentric video, spurred on by a number of relatively large, open-source data-sets, such as EPIC-KITCHENS[63], Ego4D[50], and Ego-Exo4D[64] which move away from being labelled by sets of predefined activities towards open-ended natural language descriptions.

Vision language models

Vision-language models (VLMs) are a broad class of models which process both visual, and textual data for tasks such as image-based text retrieval, image captioning, and image classification[65]. Natural language descriptions of visual content, such as alternative text descriptions of images, or summaries of video segments, are widely available on the internet, sidestepping the need for annotated data. VLMs, such as CLIP[54], and LLaVA[66], are typically trained on large data-sets of pairs of images and text, scraped from the internet, such as WebImageText[54] and LAION-5B[67], and increasingly, synthetic labels generated by frontier multimodal models, such as GPT-4, are used to make up higher quality data-sets in a secondary training stage[66]. Despite having not been explicitly trained for them, these models have shown good performance in several downstream tasks, including image classification on benchmarks such as ImageNet[68], suggesting that pretraining VLMs on large data-sets produces models which transfer well to new tasks. One recent work suggests the success of VLMs in recognising concepts in downstream tasks can be attributed to the prevalence of these concepts in their large pretraining data-sets, though with the performance scaling logarithmically with concept frequency[69].

In this work, we consider both a dual encoder VLM, CLIP[54], which quantifies the similarity between images and text, and generative VLMs, BLIP2 and LLaVA, which can be prompted to describe, and answer questions about images. All of these models have mechanisms which allow them to perform image classification in a "zero-shot" transfer setting, i.e. without having seen task-specific data, in this case, egocentric images labelled with activity intensity classes.

Methods

Our aim was to assess the performance of VLMs for predicting activity intensity classes from wearable camera images. To do this, we compared the performance of different VLMs and discriminative models on two free-living validation studies labelled with labels from the compendium of physical activity, which have known mappings to activity intensity classes.

Data processing and quality assessment

The Oxfordshire and Sichuan validation studies collected concurrent chest-worn camera (OMG Life Autographer) and wrist-worn accelerometer data (Axivity AX3). Trained human annotators reviewed the recorded sequences of images and annotated the activity depicted in each image based on the Compendium of Physical Activity[14], e.g. `occupation;interruption;13030 eating sitting`. The start and stop times of an annotation are based on the first and last image demonstrating the activity, as opposed to being based on true activity boundaries (likely between images), or fixed epochs (e.g. 1 min). It is possible to aggregate the image-timestamp-based annotations into epoch-based annotations. The estimated MET values associated with the compendium entries, along with the posture, were then used to classify the activity associated with each image as either sedentary behaviour, LIPA or MVPA, based on the provided definitions. Additional data-set and the annotation protocol details are in[19,24]. We report the median of the number of images in each intensity class per participant in Table 2, and show the spread in quartile plots in Supplementary Fig. S2b.

The images in these data-sets are egocentric, meaning there is inherent ambiguity in the participant's activities since the participants is largely unobserved. Ambiguity also arises from the low, variable frame rate (e.g. 1 image/20 s in the Oxfordshire study) and the occasional obstruction or removal of the camera. For instance, brief bursts of activity shorter than the frame-rate may be missed. All of these factors influence annotation quality. An illustration of a sequence of images captured at this frame rate is shown in Supplementary Fig. S1. In Section 2 of the Supplement, we explore the relationship between image capture rate and the number of distinguishable activities per participant, and image obscurity (darkness and variation in pixel values), against whether the image was annotated.

Images in both studies that were not labelled were excluded from the rest of our analysis. We indicate the number of labelled images in each study in Table 2, and the number of unlabelled images in each study in Supplementary Table S1. Based on the large number of unannotated images in the Sichuan data-set, we decided not to do model development on this data-set, and purely reserve it for model testing. 70% of the participants in

	Oxfordshire	Sichuan
Number of participants	161	111
Number of labelled images (% all images)	231,837 (74%)	46,184 (34%)
Median δt (1st, 3rd quartile) between images (s)	24 (23, 32)	84 (69, 88)
No. unique labels	220	110
Median instances per participant: Sedentary	884	184
LIPA	441.5	142
MVPA	81	45
Number of participants (%) aged: 0–30	45 (28%)	12 (11%)
30–50	67 (42%)	43 (41%)
50–70	39 (24%)	49 (47%)
70–100	8 (6%)	1 (1%)
Sex: Female	103 (64%)	63 (58%)
Male	58 (36%)	45 (42%)

Table 2. Summary statistics for each data-set, comparing the size, resolution and demographics between the Oxfordshire and Sichuan study. There were no reported ages for 2 participants in the Oxfordshire study. In the Sichuan study, 4 participants had no reported age, 2 had invalid ages (≥ 500), and 3 had no reported sex.

the Oxfordshire study were randomly selected for model training, 15% for validation and model selection, and 15% for testing the final models.

Eligible participants provided written informed consent prior to any study procedures taking place. Ethical approval of all experimental protocols was granted by the University of Oxford Inter-Divisional Research Ethics Committee (Ref SSD/CUREC1A/13–262) for the Oxfordshire study, and by the Sichuan Center for Disease Control and Prevention for the Sichuan study. All procedures were conducted in accordance with the Declaration of Helsinki. The egocentric images from these studies are not publicly available due to the sensitive nature of the images, but are available from the corresponding author on reasonable request. The labelled accelerometer data from the Oxfordshire study is publicly available at https://ora.ox.ac.uk/objects/uuid:99d7c092-d865-4a19-b096-cc16440cd001.

Simplifying labels

When doing exploratory data analysis, we noticed that some of the raw labels were misspelled, e.g. "office wok/computer work general", and that the same activities would be included in multiple labels with different prefixes, such as "walking;5060 shopping miscellaneous, and "5060 shopping miscellaneous". To come up with a more concise set of labels, we used a sentence embedding model[70] to embed the labels, and then used agglomerative clustering to build a dendrogram of related labels, based on their embeddings[71,72]. We then manually went through the tree, merging sets of labels with the same meaning together. We refer to this concise, semantically deduplicated set of labels as the 'clean labels'. This set of labels represents a more detailed set of colloquial activities encompassing the activities performed in the Oxfordshire study, which we use in "Generative models" section as an intermediate set of targets when predicting activity intensity.

Predicting activity intensity using computer vision

In order to asses how well computer vision methods can predict activity intensity classes from wearable cameras, we went through a process of model training, hyperparameter tuning, model selection and testing on data from unseen participants. We considered two different discriminative and three different VLMs, and for each model, we conducted a random search over the model hyperparameters[73], evaluating the performance of each hyperparameter run on the validation split. Finally, we selected the best discriminative model, and VLM, and evaluated their performance on the test split of the Oxfordshire study, and on the entire Sichuan study.

Given an image as input, the discriminative models output a vector, indicating the probability of the image belonging to one of the 3 activity intensity classes. The VLMs can further be divided into generative models, which output natural language descriptions given an image and an optional prompt as inputs, and dual-encoder models, which embed each image and a natural language description of each class into a joint embedding space, where the similarity between different images and descriptions can be quantified by looking at the similarity between their embeddings.

We investigated two generative VLMs, 3 billion parameter BLIP2[74], based on the FlanT5-XL language model[75], and 7 billion parameter LLaVA[66], and one dual-encoder model, CLIP[54]. We used the model checkpoints available on Hugging Face[76], and the exact Hugging Face model IDs are given in Supplementary Table S2. BLIP2 and LLaVA are both open-source VLMs which have shown strong performance on image captioning, with both adopting the CLIP vision encoder as a component, motivating the inclusion of CLIP as a stand-alone model to ablate the benefits of using prompted, generative VLMs, which include language models as an additional component, over a dual-encoder model.

We tested these VLMs against a commonly adapted transfer learning approach of fine-tuning a pretrained model using task specific data, and we refer to the resulting models as discriminative models. As a baseline model, we used a ResNet-50[77], pretrained on ImageNet-1K[68], and the image encoder from CLIP, pretrained on

WebImageText[54], which we refer to as ViT, which is a reference to its vision transformer architecture[78]. Though the focus of this paper is on image based classification, we also include the best sequence model found in[61], ResNet-LSTM, which has the advantage of being able to access information from multiple images.

Discriminative models

For the discriminative models, we trained the models on the training split, monitoring performance on the validation split throughout training. We used the AdamW optimizer[79] to update model weights to minimise the cross-entropy loss, and used early stopping to terminate the training, monitoring the validation cross-entropy loss, with a patience of 5. The best model found during training based on the validation loss was used to made predictions on the validation split, from which we calculated the validation metrics used to perform model selection, and study the impact of hyperparameters. For all models, we replaced the final fully connected layer of the image encoders. For the single image models, ResNet and CLIP image encoder, we replaced it with a linear layer with three outputs. The ResNet-LSTM was constructed by using a long short-term memory unit[80] to model temporal dependencies across 3 independently encoded image embeddings produced by a ResNet-50[77].

One of the most important hyperparameters for discriminative models is the learning rate[73], and for all the single-image based discriminative models we did a random search over different learning rates, batch-sizes, whether we applied data-augmentation, and whether we did full fine-tuning, or only fine-tuned the linear layer. For each model, we did 30 trials of different hyperparameters. The search space for these hyperparameters is presented in Supplementary Table S3, and the exact sweep configurations for each model are in the repository. The only hyperparameter tuning done for the ResNet-LSTM was to train three different models with learning rates, 10^{-3}, 10^{-4}, 10^{-5}. For data-augmentation, we used TrivialAugment, which samples a single augmentation uniformly at random from a set of 21 augmentations, along with a strength with which the augmentation is applied to each image[81].

Dual-encoder CLIP

As proposed in[54], we classify images by embedding them using the image encoder, and the set of labels using the text encoder. Classification is then framed as a text retrieval task where for each image, we retrieve the most similar label by looking at the cosine similarities between each image embedding, and all the label embeddings, and selecting the label associated with the largest cosine similarity.

We either used natural language descriptions of the intensity classes as targets, or used the more detailed clean labels as targets, which have a known mapping to the intensity classes. Intuitively, the set of clean labels represent more colloquial descriptions of physical activity, which may be better represented in the pretraining data-sets of VLMs compared to the intensity classes. For instance, the phrase "sedentary behaviour" might not be well represented, whereas phrases such as "lying down" which represent instances of SB, might be more prevalent. When using the intensity classes as targets, SB was represented as "sedentary behavior", LIPA as "light physical activity", and MVPA as "moderate-to-vigorous physical activity".

A similar idea of adapting pretrained VLMs by rephrasing the text targets was explored in[82], where they used a large language model to generate alternate descriptions for each of the target labels and trained a linear classifier to map between embeddings of the target labels and embeddings of the corresponding alternate descriptions. Our approach can be viewed as a non-parametric alternative to this. However, a weakness with both of these approaches is that neither of them strictly check whether an intensity class is implied by the generated description, and we show some of these failure cases in Supplementary Table S5.

Generative models

For the generative VLMs, we used different prompts to condition text generation. To evaluate whether the true intensity class could be inferred from the model's natural language description of each image, we used a text-embedding model, all-MiniLM-L12-v2, to embed the descriptions[70], and then followed a similar strategy to CLIP of mapping these descriptions to either the nearest intensity class, or the nearest clean label based on the similarity of their embeddings. In addition to varying the mapping approach, we varied the number of tokens generated, the prompt, and how we represented the activity intensity classes. We proposed an initial set of prompts, ranging from task-specific ones, e.g. "Question: What is the intensity of the physical activity in the image? Options: Sedentary, Light, Moderate-Vigorous. Short answer:", to more generic descriptive prompts, e.g. "a photo of". We also augmented the set of prompts by asking proprietary large language models, ChatGPT, Claude, and Gemini, to suggest similar prompts and selecting sensible ones. The final set of 17 prompts is included in the repository. The exact hyperparameters that were varied for each model are shown in Supplementary Table S2.

Evaluation

We assessed each model's performance across activity intensity classes using Cohen's κ score, and the performance per class using the F_1-score of the class[72]. The Cohen's κ score (κ or "kappa" for short in Figures) is 0 if the model's performance is on par with a random classifier, and 1 if all instances were correctly predicted. The F_1-score for a class is the harmonic mean of the recall, the proportion of instances of the class that were correctly predicted, and the precision, the proportion of predictions of that class that were correct. Since there is a class imbalance, reporting per class F1-scores helps avoid inflating the performance of classifiers that are biased towards predicting the majority class. We calculated these metrics per participant and present the spread of the per-participant scores in our results. This does however come with the caveat that some participants had relatively few instances of LIPA and MVPA, thus the estimate of these metrics at the participant level had high variance.

Results

In "Data processing and EDA" section, we present the results from data-processing and exploratory data analysis, highlighting some of the challenges of modelling free-living egocentric timelapses, and in "Model results" section, we present results from model selection, motivating the choice of the best models. Finally, we present the performance of the best vision-language and discriminative model.

Data processing and EDA

The Oxfordshire study had 231,837 (from an original 312,585) images with non-trivial labels from 161 participants (Table 2), i.e not labelled as "uncodeable", or "undefined". The median time interval, $_{\delta t}$, (1st, 3rd quartile) between images was 24 s (23, 32). The Sichuan study had a much larger median time interval of 84 s (69, 88), and a much smaller proportion of images with non-trivial labels of 46,184 images (from an original 132,850 images) from 111 participants.

We estimated the time covered in each study as

$$\text{Time covered (h)} = \frac{\text{No. labelled images} \times \text{median } \delta t \text{ between images (s)}}{60 \times 60},$$

suggesting that there were 1546 h of labelled data in the Oxfordshire study and 1078 h of labelled data in the Sichuan study, though this is an overestimate because the low temporal resolution, particularly in the Sichuan study, means that knowing the activity in each image does not necessarily mean we continue to know the activity in an 84-s window surrounding that image.

One noticeable feature of both data-sets is the large number of images that were difficult to label. We differentiate between images that were unlabelled, and images where the labels were unknown, which includes both unlabelled images, and images with labels such as "image dark/blurred/obscured". Although the number of unlabelled images in both study was relatively low (7.57% for the Oxfordshire study and 1.31% for the Sichuan study), the number of images with unknown labels was very high (25.8% for the Oxfordshire study and 65.2% for the Sichuan study).

The median δt between frames was much lower in the Sichuan study, compared to the Oxfordshire study. Supplementary Fig. S2a, echoes this, though by showing the median δt for each participant, also reveals that participants clustered around four distinct median capture rates, suggesting that different base capture rates were erroneously set on the Autographers, leading to these different resolutions. Although the estimated number of hours captured in each study are of similar orders of magnitude, the number of annotated events in the Sichuan study is much lower, pointing to the lower capture rate set on the devices as being a bottleneck for the resolution of the annotations.

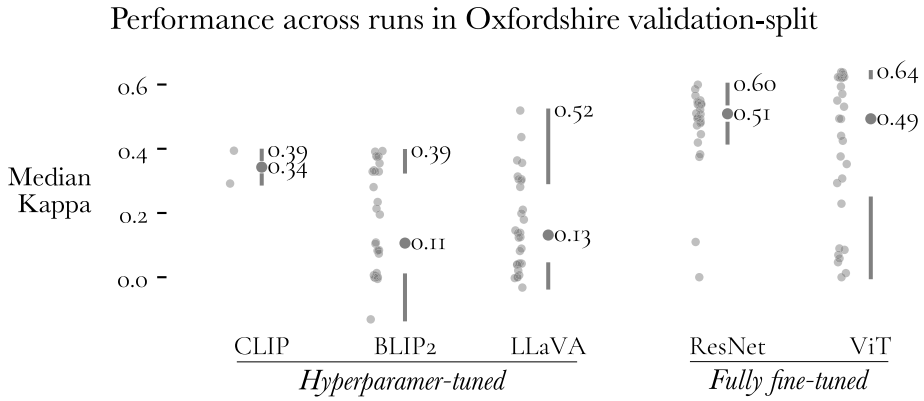
Model results

We used the model's validation performance on the Oxfordshire study to identify promising models, and for each model, promising hyperparameters. The left side of Fig. 2a shows that for the VLMs, differences in the prompts, mapping approach, and number of generated tokens resulted in large differences in validation performance (κ scores range from 0 to 0.5). The right side of Fig. 2a shows the validation performance of fine-tuned DMs, which tended to be better than the VLMs, though also displays a sensitivity to different hyperparameters.

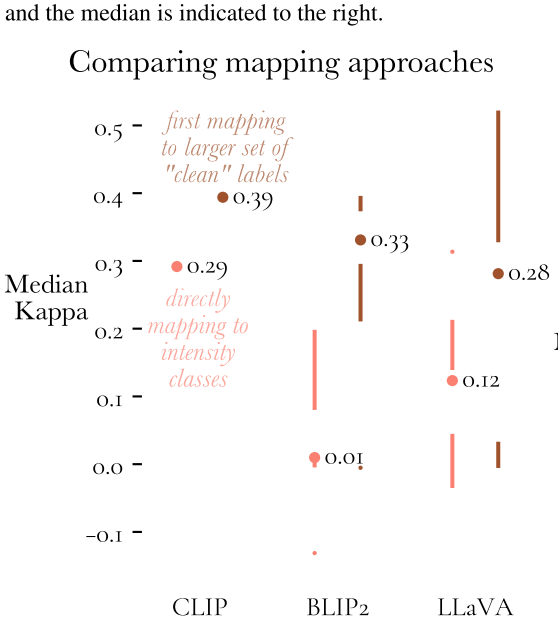
For the VLMs, we highlight the mapping approach as one of the hyperparameters associated with this variation. Figure 2b visualises the difference in performance between runs that used the larger-set of more colloquial activities as targets and those which directly used SB, LIPA, and MVPA as targets. Across all VLMs, the median performance of the runs that adopted the more colloquial targets was higher. Despite this, the best performing VLM, LLaVA, which was prompted, "Walking, Running, Sitting, Standing, Other. Based on the objects in the image, what is the person likely doing?", had its responses directly mapped to one of the activity classes, and not the clean labels.

Examining the spread in validation performance across different hyperparameter runs for the ResNet and ViT in isolation suggests that the ResNet is the more robust model, since the median of the median κ scores is higher, and the interquartile range is narrower. Figure 2 celaborates on this picture, revealing that the combination of doing full fine-tuning and using a high learning rate ($l \geq 10^{-4}$) was particularly detrimental for the ViT, and that when when only fine-tuning the last layer, the performance of the ViT was consistently better than the performance of the ResNet. We saw better performance from fine-tuning the last layer as opposed to full fine-tuning, despite the latter being a more flexible model adaptation technique. In general, lower learning rates were associated with better validation performance, with the relationship between the logarithm of the learning rate and the median κ roughly following a negative linear line, suggesting that performance could be further improved by using even lower learning rates.

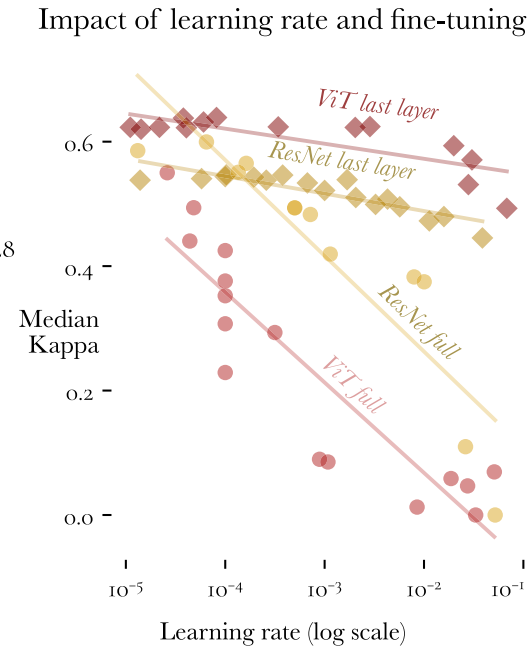
Finally, we selected the best performing vision-language (LLaVA), and discriminative model (ViT), and assessed their performance on the withheld test-set (Fig. 1). SB in the Oxfordshire test-set was well predicted by all models, with median F_1-scores of 0.89 (0.84, 0.92) for LLaVA and 0.91 (0.86, 0.95) for ViT. Predictive performance on LIPA and MVPA, although much better than chance performance, was worse than SB, which a median F_1-score of 0.60 (0.56, 0.67) for LLaVA, and 0.70 (0.63, 0.79)) and for ViT. The spread in the performance across participants was large for these behaviours, particularly MVPA. We found a large drop in performance when going from the Oxfordshire study, where models were trained and/or hyperparameter-tuned, to the Sichuan study. The largest drop in performance was for the ViT, which went from a median κ of 0.67 (0.60, 0.74), which can be interpreted as showing substantial agreement relative to the human annotations[83], to 0.19 (0.10, 0.30), which only shows fair agreement. For LLaVA the drop in performance was from a median κ of 0.54 (0.64, 0.49) to 0.26 (0.15, 0.37).



(a) Quartile plots show the range of median κs on the Oxfordshire validation-split across the 30 runs for each model (except for CLIP). Each run randomly sampled a different set of hyperparameters. The median κ of each run is shown as a jittered column of dots to the left of each quartile plot. The maximum of the median κs is indicted above the quartile plot, indicating the performance of the best found hyperparameters for each model, and the median is indicated to the right.



(b) Quartile plots comparing directly calculating the similarity between the generative model image descriptions, or image embeddings for CLIP, and the intensity labels, versus calculating their similarity to a broader set of activities with known mappings to the activity labels. The results reflect the spread of the median κ across runs with different randomly sampled prompts, and number of tokens generated.



(c) Scatter plot showing the median κ of runs with different learning rates, and where either only the last layer, or the full model was fine-tuned. Median κs were higher for ViT than ResNet runs when only the last layer was fine-tuned, and considerably worse when fine-tuning the full model.

Fig. 2. Impact of different hyperparameters on the performance of each model on the validation-set of the Oxfordshire study.

Whereas human annotators were allowed to view the entire history of a participant's day when annotating each image, these models make predictions based on single images. In order to estimate human performance in the same setting, one of the present authors manually labelled > 500 randomly selected images from the test-set of each study, without temporal context, and obtained a median κ of 0.63 (0.45, 0.72) on the Oxfordshire study, and 0.572 (0.46–0.61) on the Sichuan study. The performance on the Oxfordshire study is similar to the performance observed for the best model, though noticeably better than the model performance of the Sichuan study.

Though not strictly a fair comparison to the single-image models, we also tested the performance of a sequential model (ResNet-LSTM) to investigate the benefits of going beyond single frame predictions. This model consistently had similar or slightly better F_1-scores for each of the activity intensity classes compared to the best single-image model, and obtained a median κ of 0.66 (0.59, 0.72) on the Oxfordshire study and median

κ of 0.31 (0.18, 0.41) on the Sichuan study suggesting that performance can be further improved by developing sequential models for these sparse sequences of images.

Finally, if we look at the accuracy of the models, which is misleading in that it is dominated by performance on the majority class, but relevant in that it relates to the fraction of images that would have to be corrected by a human annotator, both of the best models achieve an accuracy $> 80\%$ on the Oxfordshire test-set, and $> 50\%$ on the Sichuan study.

Discussion

We compared the performance of VLMs and DMs on predicting activity intensity in two free-living validation studies, and found that SB was well predicted in unseen participants within the Oxfordshire study, but that LIPA and MVPA were less well predicted, and all models generalised poorly to the Sichuan study. The overall accuracy of the models on unseen participants in the Oxfordshire study suggest they might still be useful for labelling wearable camera images, especially in free-living data where SB typically makes up the majority of instances as seen in Table 2, though within similar studies to ones they have been adapted for.

Similar work by[84], though based on third-person still frames from a GoPro, found that their best model at distinguishing between SB, light, moderate, and vigorous intensity physical activity, a tree-based model (XGBoost[85]) based on features from AlphaPose[86], was able to do so with an accuracy of 68.6%. Although they separate out moderate and vigorous physical activity into distinct classes, we can calculate performance metrics compatible with this work by combining the rows and columns for these classes in the confusion matrix in Table 3 of their work, included here in Supplementary Table S4, comparing it to the confusion matrix in Supplementary Fig. S3.

The overall accuracy for predicting activity intensity of XGBoost was 69.2%, compared to the finetuned ViT in this work, which achieved an accuracy of 84.6% on unseen data in the Oxfordshire study, and LLaVA, which achieved an accuracy of 80.9%. The improved performance of ViT and LLaVA in this context is in part driven by better recall of SB, which was predicted with a recall of 71.6% in[15], but with recalls of 90.7% and 89.1% for ViT and LLaVA, respectively, in this work, and there was also a higher proportion of SB in studies used in this work, thus the accuracy was more heavily weighted by SB. If we consider the average of the per-class recalls, which weights the classes equally, the performance is closer, 70.0% for XGBoost, 76.8% for ViT and 72.5% for LLaVA.

However, there are many limitations to this comparison, including the varying perspectives (first vs. third person), and frame-rates (0.05 vs. 30 fps) with which each study captured footage. Annotating activity intensity classes from third person video recordings is a more accurate way of validating device-measured activity intensity measurements[10]. Martinez et al.[87] compared using sparse sequences of images captured by wearable cameras to assess posture against the activPAL and reported that, although the bias in estimates of sitting time was not significant, there was significant bias in estimates of standing and movement time. On the other hand, the use of egocentric cameras for capturing validation data is more scalable since it does not require researchers to follow participants, enabling the Oxfordshire and Sichuan validation studies to collect data from 100+ participants each.

Supplementary Fig. S2c highlights the challenge of interpreting images in poorly lit conditions, with a large number of dark images left unannotated. Consistent with this, Supplementary Table S6 shows that both LLaVA and the ViT performed worse in the darkest 5% of images (LLaVA median κ: 0.31 [0.12, 0.44]; ViT: 0.33 [0.18, 0.55]) compared to brighter images in the Oxfordshire test-set. This highlights the broader issue that low-visibility conditions, frequently encountered with wearable cameras in free-living settings, substantially limit annotation quality, whether human- or model-derived. 26% of images in the Oxfordshire dataset remained unannotated by humans, likely due in part to low visibility. Consistent with this, both visual-language models demonstrated notably reduced performance on the darkest 5% of images. While lighting clearly impacts annotation reliability, the exact proportion of annotation loss attributable specifically to low visibility remains uncertain, especially given the higher proportion of unannotated data in the Sichuan data-set (66%), where additional factors such as its much lower capture rate are likely influential.

The focus on models based on single images was motivated by the availability of VLMs in this setting, and the lack of models for sparse sequences of images. However, predicting activities from single images is a notable obstacle, and our limited analysis of one annotator's performance in this regime suggests that the current levels of performance on the Oxfordshire study are close to human performance based on single images. Beyond single-image models, the ResNet-LSTM, performed slightly better than the single-image models, and did not undergo hyperparameter tuning to the same extent. This suggests the necessity of moving beyond single-frame models, and developing and assessing multi-modal models which can handle sparse sequences of images.

A sentence embedding model was used to embed model responses from off-the-shelf VLMS so that we could quantify their similarity to activity intensity classes. However, this introduced some semantic mismatches where model responses were mapped to activity classes which were not implied by the response (Supplementary Table S5). These VLMs could be further improved by adaptation techniques such as parameter efficient fine-tuning[88], or prompt engineering[89]. This work examined performance in two populations of ambulant adults, and may not reflect performance in other populations, such as non-ambulant people. This was an imbalanced problem, and we observed high variation in the performance estimates of the less prevalent classes. Our performance estimates could have been more robust by adopting methods such as cross-validation, though at the expense of these experiments being more computationally expensive. Each hyperparameter-tuning run took an average of 5 h to complete on a V100 GPU for the ResNet, the smallest model.

Despite these limitations, this work was able to assess performance in studies collected in free-living conditions in a large number of participants revelative to existing wearable validation studies, and it assessed generalisation using an independently collected study. Activity intensity classes have been adopted in a number of downstream epidemiological works[6,16,17], and we used definitions compatible with this field of research. The

application of VLMs to estimating activity intensity is novel, and also raises the possibility of measuring new behaviours, such as environmental exposures, social interactions, eating and drinking behaviours, without the need for task specific training. An application using VLMs to label outdoor time to validate wrist-worn light sensors is concurrently being explored.

Improvements in technology not only suggest new ways of analysing validation studies, but also conducting them. Tran et al.[90] proposed developing wearable cameras which cost less, and Mamish et al.[91] proposed a wearable camera able to capture footage at high frame-rates while lasting several days. Commercially available body cameras, such as those manufactured by BOBLOV and MIUFLY, are commercially available and able to record 15 h of video footage on a single charge. The adoption of these cameras in future validation studies would reduce the annotation uncertainty due to low frame-rates whilst making it easier to adopt activity recognition approaches developed for egocentric video[92]. Although we focus on wearable cameras as a way of informing ground truth labels to validate and train measurement approaches typically using other wearable sensors, wearable cameras have also been used in small health studies[32,93,94] as the measurement device themselves. Given the range of behaviours that can be measured simultaneously from a single camera in comparison to other wearables, and the human interpretable nature of the modality, one might be tempted to directly adopt them in health studies. However, the large amount of information captured by these cameras raises various ethical issues, and has made it unlikely that they will be adopted for large scale health studies[20,95,96].

Although we have made the distinction between the broader field of activity recognition and recognising health relevant activity intensity classes, progress in the former is vital to this task, and should not be disregarded. This work showed that the performance of generalist VLMs is similar to domain specific discriminative models, and progress on developing more capable generalist models might well outpace approaches reliant on annotated wearable data. This suggests the importance of exploring similarities between more mainstream computer vision research and the present study. There is also additional work needed in applying methods from fields such as continual learning, active learning and uncertainty quantification so that models can be adapted and assessed 'on the fly' to efficiently learn from new labelled data, so that human input can be used efficiently in correcting the most informative instances, and so that models can indicate which samples they cannot reliably label. After all, model accuracy is only one aspect impacting the efficiency of labelling wearable data-sets.

Conclusions

In this paper we assessed the performance of fine-tuned discriminative models and vision-language models on the simple, but important task of predicting activity intensity classes from two free-living validation studies, each comprising over 100 participants, conducted in Oxfordshire, UK, and Sichuan, China. Sedentary behaviour was well predicted within unseen participants from a seen population by both types of models. Random searches over different hyperparameters revealed the importance of how activity intensity classes were phrased when using vision-language models, and the importance of minimal fine-tuning for the discriminative models. Although none of these approaches pass the threshold required for trained human annotators, we only focused on activity prediction based on single images, which is a notable handicap on model performance, and initial results reproducing a sequence-based classifier in this setting shows slightly better performance. Although several times bigger than existing validation studies, the studies used here were still prone to errors in the ground-truth labels arising from the sparsity of the images, and large numbers of obscure images. Despite these limitations, we would recommend the adoption of the best models found in this study to label sedentary behaviour in free-living studies as they are freely available, relatively easy to adapt and can substantially reduce the annotation burden given the prevalence of sedentary behaviour. We would also encourage research groups conducting wearable camera based validation studies to consider moving to newer wearable cameras which are able to record videos for the full waking day, which would significantly lower the uncertainty in the ground-truth labels of physical activity.

Data availability

The egocentric images from these studies are not publicly available due to the sensitive nature of the images, but are available from the corresponding author on reasonable request. The labelled accelerometer data from the Oxfordshire study is publicly available at https://ora.ox.ac.uk/objects/uuid:99d7c092-d865-4a19-b096-cc16440cd001. Code available at https://github.com/oxwearables.

Received: 27 May 2025; Accepted: 19 September 2025
Published online: 24 October 2025

References

1. Wasfy, M. M. & Lee, I.-M. Examining the dose–response relationship between physical activity and health outcomes. *NEJM Evid.* **1**(12), EVIDra2200190 (2022).
2. Servais, L. et al. First regulatory qualification of a digital primary endpoint to measure treatment efficacy in DMD. *Nat. Med.* **29**(10), 2391–2392 (2023).
3. Troiano, R. P., Stamatakis, E. & Bull, F. C. How can global physical activity surveillance adapt to evolving physical activity guidelines? Needs, challenges and future directions. *Br. J. Sports Med.* **54**(24), 1468–1473 (2020).
4. Logacjov, A., Herland, S., Ustad, A. & Bach, K. SelfPAB: Large-scale pre-training on accelerometer data for human activity recognition. *Appl. Intell.* **54**(6), 4545–4563 (2024).
5. Yuan, H. et al. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ Digit. Med.* **7**(1), 91 (2024).
6. Walmsley, R. et al. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. *Br. J. Sports Med.* **56**(18), 1008–1017 (2022).

7. Willetts, M., Hollowell, S., Aslett, L., Holmes, C. & Doherty, A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Sci. Rep.* **8**(1), 7961 (2018).
8. Doherty, A. et al. Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLoS ONE* **12**(2), e0169649 (2017).
9. Bao, L. & Intille, S. S. Activity recognition from user-annotated acceleration data. In *International Conference on Pervasive Computing*, 1–17 (Springer, 2004).
10. Keadle, S. K., Lyden, K. A., Strath, S. J., Staudenmayer, J. W. & Freedson, P. S. A framework to evaluate devices that assess physical behavior. *Exerc. Sport Sci. Rev.* **47**(4), 206–214 (2019).
11. Thomaz, E. & Dimiccoli, M. Acquisition and analysis of camera sensor data (lifelogging). In *Mobile Sensing in Psychology: Methods and Applications*, 277 (2023).
12. Tufte, E. R. *The Visual Display of Quantitative Information* 2nd edn. (Graphics Press, 2002).
13. Tremblay, M. S. et al. Sedentary behavior research network (SBRN)-terminology consensus project process and outcome. *Int. J. Behav. Nutr. Phys. Act.* **14**, 1–17 (2017).
14. Ainsworth, B. E. et al. 2011 compendium of physical activities: A second update of codes and met values. *Med. Sci. Sports Exerc.* **43**(8), 1575–1581 (2011).
15. Keadle, S. K. et al. Using computer vision to annotate video-recoded direct observation of physical behavior. *Sensors* **24**(7), 2359 (2024).
16. Schalkamp, A.-K., Peall, K. J., Harrison, N. A. & Sandor, C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. *Nat. Med.* **29**(8), 2048–2056 (2023).
17. Shreves, A. H., Small, S. R., Travis, R. C., Matthews, C. E. & Doherty, A. Dose–response of accelerometer-measured physical activity, step count, and cancer risk in the UK Biobank: A prospective cohort analysis. *Lancet* **402**, S83 (2023).
18. Bull, F. C. et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *Br. J. Sports Med.* **54**(24), 1451–1462 (2020).
19. Chan, S. et al. Capture-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Sci. Data* **11**(1), 1135 (2024).
20. Kelly, P. et al. An ethical framework for automated, wearable cameras in health behavior research. *Am. J. Prev. Med.* **44**(3), 314–319 (2013).
21. Ainsworth, B. E., Herrmann, S. D., Jacobs Jr, D. R., Whitt-Glover, M. C. & Tudor-Locke, C. A brief history of the compendium of physical activities. *J. Sport Health Sci.* **13**(1), 3 (2024).
22. Bureau of Labor Statistics. *American Time Use Survey, 2024*. Accessed 13 May 2024.
23. Herath, S., Harandi, M. & Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **60**, 4–21 (2017).
24. Chen, Y. et al. Device-measured movement behaviours in over 20,000 China Kadoorie Biobank participants. *Int. J. Behav. Nutr. Phys. Act.* **20**(1), 138 (2023).
25. Byrne, N. M., Hills, A. P., Hunter, G. R., Weinsier, R. L. & Schutz, Y. Metabolic equivalent: One size does not fit all. *J. Appl. Physiol.* **99**, 1112–1119 (2005).
26. Walmsley, R. *Device-Measured 24-Hour Movement Behaviours and Risk of Incident Cardiovascular Disease*. PhD thesis, University of Oxford (2022).
27. Kozey, S. L., Lyden, K., Howe, C. A., Staudenmayer, J. W. & Freedson, P. S. Accelerometer output and MET values of common physical activities. *Med. Sci. Sports Exerc.* **42**(9), 1776 (2010).
28. Pober, D. M., Staudenmayer, J., Raphael, C. & Freedson, P. S. Development of novel techniques to classify physical activity mode using accelerometers. *Med. Sci. Sports Exerc.* **38**(9), 1626 (2006).
29. Montoye, A. H. K., Begum, M., Henning, Z. & Pfeiffer, K. A. Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiol. Meas.* **38**(2), 343–357 (2017).
30. Hills, A. P., Mokhtar, N. & Byrne, N. M. Assessment of physical activity and energy expenditure: An overview of objective measures. *Front. Nutr.* **1**, 5 (2014).
31. Kim, Y., Barry, V. W. & Kang, M. Validation of the ActiGraph GT3X and activPAL accelerometers for the assessment of sedentary behavior. *Meas. Phys. Educ. Exerc. Sci.* **19**(3), 125–137. https://doi.org/10.1080/1091367X.2015.1054390 (2015).
32. Kerr, J. et al. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. *Am. J. Prev. Med.* **44**(3), 290–296 (2013).
33. Chasan-Taber, L. et al. Update and novel validation of a pregnancy physical activity questionnaire. *Am. J. Epidemiol.* **192**(10), 1743–1753 (2023).
34. Nawab, K. A. et al. Accelerometer-measured physical activity and functional behaviours among people on dialysis. *Clin. Kidney J.* **14**(3), 950–958 (2021).
35. Martinez, J. *Accuracy and Precision of Wearable Camera Media Annotations to Estimate Dimensions of Physical Activity and Sedentary Behavior*. PhD thesis, University of Wisconsin-Milwaukee (2024).
36. Giurgiu, M. et al. Quality evaluation of free-living validation studies for the assessment of 24-hour physical behavior in adults via wearables: Systematic review. *JMIR mHealth uHealth* **10**(6), e36377 (2022).
37. Femiano, R., Werner, C., Wilhelm, M. & Eser, P. Validation of open-source step-counting algorithms for wrist-worn tri-axial accelerometers in cardiovascular patients. *Gait Posture* **92**, 206–211 (2022).
38. Alphen, H. J. M., Waninge, A., Minnaert, A. E. M. G., Post, W. J. & Putten, A. A. J. Construct validity of the Actiwatch-2 for assessing movement in people with profound intellectual and multiple disabilities. *J. Appl. Res. Intell. Disabil.* **34**(1), 99–110 (2021).
39. Bach, K. et al. A machine learning classifier for detection of physical activity types and postures during free-living. *J. Meas. Phys. Behav.* **5**(1), 24–31 (2021).
40. Marcotte, R. T. et al. Estimating sedentary time from a hip- and wrist-worn accelerometer. *Med. Sci. Sports Exerc.* **52**(1), 225 (2020).
41. Koenders, N. et al. Validation of a wireless patch sensor to monitor mobility tested in both an experimental and a hospital setup: A cross-sectional study. *PLoS ONE* **13**(10), e0206304 (2018).
42. Gershuny, J. et al. Testing self-report time-use diaries against objective instruments in real time. *Sociol. Methodol.* **50**(1), 318–349 (2020).
43. Doherty, A. et al. GWAS identifies 14 loci for device-measured physical activity and sleep duration. *Nat. Commun.* **9**(1), 1–8 (2018).
44. Mann, S. Wearable computing: A first step toward personal imaging. *Computer* **30**(2), 25–32 (1997).
45. Aizawa, K., Ishijima, K. & Shiina, M. Summarizing wearable video. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, Vol. 3, 398–401 (IEEE, 2001).
46. Bush, V. et al. As we may think. *Atl. Mon.* **176**(1), 101–108 (1945).
47. Feichtenhofer, C., Fan, H., Malik, J. & He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6202–6211 (2019).
48. Zhang, C.-L., Wu, J. & Li, Y. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510 (Springer, 2022).
49. Momeni, L., Caron, M., Nagrani, A., Zisserman, A. & Schmid, C. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15579–15591 (2023).

50. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012 (2022).
51. Lin, K. Q. et al. Egocentric video-language pretraining. *Adv. Neural Inf. Process. Syst.* **35**, 7575–7586 (2022).
52. Pramanick, S., Song, Y., Nag, S., Lin, K. Q., Shah, H., Shou, M. Z., Chellappa, R. & Zhang, P. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5285–5297 (2023).
53. Bock, M., Van Laerhoven, K. & Moeller, M. Weak-annotation of HAR datasets using vision foundation models. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, ISWC '24, 55–62 (Association for Computing Machinery, New York, NY, USA, 2024).
54. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763 (PMLR, 2021).
55. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., 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* (2024). arXiv:2304.07193 [cs].
56. Carreira, J. & Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308 (2017).
57. Wang, P. & Smeaton, A. F. Using visual lifelogs to automatically characterize everyday activities. *Inf. Sci.* **230**, 147–161 (2013).
58. Moghimi, M., Wu, W., Chen, J., Godbole, S., Marshall, S., Kerr, J., & Belongie, S. Analyzing sedentary behavior in life-logging images. In *2014 IEEE International Conference on Image Processing (ICIP)*, 1011–1015 (IEEE, 2014).
59. Castro, D., Hickson, S., Bettadapura, V., Thomaz, E., Abowd, G., Christensen, H., & Essa, I. Predicting daily activities from egocentric images using deep learning. In *proceedings of the 2015 ACM International symposium on Wearable Computers*, 75–82 (2015).
60. Cartas, A., Marín, J., Radeva, P. & Dimiccoli, M. Recognizing activities of daily living from egocentric images. In *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20–23, 2017, Proceedings 8*, 87–95 (Springer, 2017).
61. Cartas, A., Radeva, P. & Dimiccoli, M. Activities of daily living monitoring via a wearable camera: Toward real-world applications. *IEEE Access* **8**, 77344–77363 (2020).
62. Cartas, A., Talavera, E., Radeva, P., & Dimiccoli, M. Understanding event boundaries for egocentric activity recognition from photo-streams. In *International Conference on Pattern Recognition*, 334–347 (Springer, 2021).
63. Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *Int. J. Comput. Vis.*, pp. 1–23 (2022).
64. Grauman, K. et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400 (2024).
65. Li, C. et al. Multimodal foundation models: From specialists to general-purpose assistants. *Found. Trends Comput. Graph. Vis.* **16**(1–2), 1–214 (2024).
66. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892–34916 (2024).
67. Schuhmann, C. et al. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).
68. Deng, J. et al. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).
69. Udandarao, V. et al. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint* arXiv:2404.04125 (2024).
70. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using Siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 (2019).
71. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Vol. 2 (Springer, 2009).
72. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
73. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
74. Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint* arXiv:2301.12597 (2023).
75. Chung, H. W. et al. Scaling instruction-finetuned language models. *arXiv preprint* arXiv:2210.11416 (2022).
76. Wolf, T. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint* arXiv:1910.03771 (2019).
77. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
78. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 [cs] (2021).
79. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019).
80. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997).
81. Muller, S. G., & Hutter, F. TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 754–762 (IEEE, Montreal, 2021).
82. Mirza, M. J. et al. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *Scjefie* **10**, 10 (2023).
83. Richard Landis, J. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159 (1977).
84. Keadle, S. K. et al. Evaluation of within-and between-site agreement for direct observation of physical behavior across four research groups. *J. Meas. Phys. Behav.* **1**(aop), 1–9 (2023).
85. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
86. Fang, H.-S. et al. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(6), 7157–7173 (2022).
87. Martinez, J. et al. Validation of wearable camera still images to assess posture in free-living conditions. *J. Meas. Phys. Behav.* **4**, 47–52 (2021).
88. Wang, L. et al. Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artif. Intell. Rev.* **58**(8), 227 (2025).
89. Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V. & Torr, P. A systematic survey of prompt engineering on vision-language foundation models. arXiv:2307.12980 [cs] (2023).
90. Tran, Q.-Li., Nguyen, B., Jones, G. J. F. & Gurrin, C. Memorilens: A low-cost lifelog camera using raspberry pi zero. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1255–1259 (2024).
91. Mamish, John et al. Nir-sighted: A programmable streaming architecture for low-energy human-centric vision applications. *ACM Trans. Embedd. Comput. Syst.* **23**, 1–26 (2024).
92. Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., Huang, Y., Wang, Y., Lu, T., Wang, L. & Qiao, Y. EgoVideo: Exploring egocentric foundation model and downstream adaptation. arXiv:2406.18070 [cs] (2024).

93. Doherty, A. R. et al. Use of wearable cameras to assess population physical activity behaviours: An observational study. *Lancet* **380**, S35 (2012).
94. Gage, R. et al. Fun, food and friends: A wearable camera analysis of children's school journeys. *J. Transp. Health* **30**, 101604 (2023).
95. Mok, T. M., Cornish, F. & Tarr, J. Too much information: Visual research ethics in the age of wearable cameras. *Integr. Psychol. Behav. Sci.* **49**, 309–322 (2015).
96. Meyer, L. E. et al. Using wearable cameras to investigate health-related daily life experiences: A literature review of precautions and risks in empirical studies. *Res. Ethics* **18**(1), 64–83 (2022).

Acknowledgements
Thank you to Shing Chang, Hang Yuan, Aidan Acquah, Laura Brocklebank, Jerred Chen and Freddie Bickford Smith for valuable advice over the course of this project. We are grateful to Huaidong Du for facilitating access to the Sichuan validation study, and we extend our thanks to all those involved in the collection and annotation of the validation datasets. Finally, we are grateful to the participants for their willingness to participate in these studies.

Author contributions
A.S. led the study design, data analysis, and drafting of the manuscript, and contributed to project conceptualization. R.C. contributed to project conceptualization, provided supervision, and reviewed and suggested edits to the manuscript. B.M. provided supervision, offered technical guidance, and reviewed and suggested edits to the manuscript. X.C. contributed to data collection. A.D. contributed to project conceptualization, supervised the study, and reviewed and suggested edits to the manuscript. All authors read and approved the final version.

Funding
Abram Schönfeldt is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). Aiden Doherty's research team is supported by a range of grants from the Wellcome Trust [223100/Z/21/Z, 227093/Z/23/Z], Novo Nordisk, Swiss Re, Boehringer Ingelheim, National Institutes of Health's Oxford Cambridge Scholars Program, EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1), British Heart Foundation Centre of Research Excellence (grant number RE/18/3/34214), and funding administered by the Danish National Research Foundation in support of the Pioneer Centre for SMARTbiomed. Xiaofang Chen acknowledges support from the Noncommunicable Chronic Diseases–National Science and Technology Major Project (2023ZD0510100) and the National Natural Science Foundation of China (82192900, 82192901, 82192904, 81390540, 91846303). For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-21350-6.

Correspondence and requests for materials should be addressed to A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.