Towards Practical, Best Practice Video Annotation to Support Human Activity Recognition

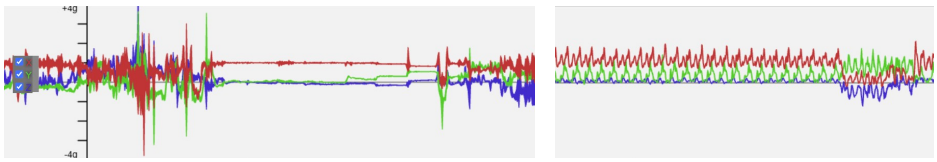
Hoan Tran✉[1] [0000−0003−4888−0513],
Veronika Potter[1] [0009−0005−7256−7713], Umberto Mazzucchelli[1] [0009−0000−1071−6027],
Dinesh John[1,2] [0009−0002−7095−1023], and Stephen Intille[1] [0000−0002−0287−2553]

[1] Northeastern University, Boston MA 02115, USA
{tran.hoan1,potter.v,mazzucchelli.u,s.intille}@northeastern.edu
[2] dineshjohn@yahoo.com

Abstract. Researchers need ground-truth activity annotations to train and evaluate wearable-sensor-based activity recognition models. Oftentimes, researchers establish ground truth by annotating the video recorded while someone engages in activity wearing sensors. The "gold-standard" video annotation practice requires two trained annotators independently annotating the same footage with a third domain expert resolving disagreements. Because such annotation is laborious, widely-used datasets have often been annotated using only a single annotator per video. Because the research community is moving towards collecting data of more complex behaviors from free-living people 24/7 and annotating more granular, fleeting activities, the annotation task grows even more challenging; the single-annotator approach may yield inaccuracies. We investigated a "silver-standard" approach: rather than using two independent annotation passes, a second annotator *revises* the work of the first annotator. The proposed approach reduced the total annotation time by 33% compared to the gold-standard approach, with near-equivalent annotation quality. The silver-standard label was in higher agreement with the gold-standard label than the single-annotator label, with Cohen's κ of 0.77 and 0.68 respectively on a 16.4 h video. The silver-standard labels also had higher inter-rater reliability than the single-annotator labels, with the respective mean Cohen's κ across six videos (92 h of total footage) of 0.79 and 0.68.

Keywords: Video · Annotation · Taxonomy · Human Activity Recognition

1 Motivation



(a) A (about 2 min) snippet of CAPTURE-24 [4] where the ground-truth label is *walking* [16].

(b) A (about 1 min) snippet of *walking* activity in our to-be-released dataset.

Fig. 1. Examples of wrist-worn accelerometer signals from two different free-living datasets with the same *walking* label, but the actual accelerometer signals differ. Non-movement (flat-line) in Figure 1a suggests *non-walking*, meaning the provided ground-truth may be inaccurate.

Accurate measurement of physical activity and other everyday behavior supports research in many fields. An "objective" and non-burdensome approach is to use wearable

motion sensors such as accelerometers to continuously measure behaviors [13]. This approach has been used to measure sedentary and physical activity behaviors in national health surveillance studies [13, 5], and research is ongoing in the field of human activity recognition (HAR) to develop models that can detect not only activity levels, but also specific types of activities with high fidelity. To do so, researchers must train and validate machine learning models with annotated activity datasets with wearable sensors where accurate ground-truth labels of physical activity have been obtained. Annotation of video where people are filmed performing activities of interest is considered the best approach to establish such reliable labels [7]. Studies to date have been mostly limited to small amounts of annotated data with simple activity taxonomies (i.e., a small set of possible labels for activities). Scaling up algorithm training and validation to improve algorithm performance, however, will require labeling larger amounts of video data (using more complex activity taxonomies) as people engage in unconstrained, free-living activity. This is an arduous task.

When labeling video for research, the "gold-standard" approach is to have two trained annotators independently label each video segment, and then to have a third trained domain expert resolve disagreements [18, 17]. While manageable for a few hours of video, this approach does not scale to larger annotation tasks required for training and evaluating robust machine learning models. In practice, for HAR datasets, often only one trained annotator labels each video segment [17]. An alternative method that researchers have explored to control cost and increase speed is crowd-sourced annotation [2], but video footage obtained from free-living volunteers in research studies is often private and thus not suitable for crowd-sourcing.

Not only are researchers interested in compiling datasets of free-living activity from larger cohorts for longer periods of time, they also desire richer activity taxonomies to be labeled to develop and evaluate future state-of-the-art algorithms for detecting novel activities. When labeling video footage acquired from the laboratory, annotators are likely aware of the entire small set of possible activities and the timing of all activities performed in the video. Additionally, lab-controlled data collection allows for a dedicated assistant to film the entire data collection session from a third-person perspective. In contrast, labeling activity of people in unconstrained free-living settings will generally require a larger activity taxonomy and labeling video acquired from a front-facing camera that might not adequately capture all activity details, leading to poor annotation quality [1]. Poor-quality annotations, including incorrect labels (Figure 1), missing labels, or labels with imprecise start/stop times, can invalidate HAR evaluation results and stunt research.

In our own work, we are confronting the daunting task of labeling 22,000 hours of free-living behavior from front-facing video at the second-by-second level using an activity/posture taxonomy including 49 unique activity types and 13 different postures. Motivated by this task, we investigated the merits of the gold-standard annotation approach, single-annotator approach, and the proposed "silver-standard approach" that could be used to reduce labor and related costs relative to the gold standard method and improve annotation over a single annotator. Based on our annotation experiments described in the remainder of this paper and our ongoing efforts to label our large dataset, we demonstrate the proposed silver-standard protocol can:

- Reduce the person-hours required for annotation by 33% compared to the gold-standard approach and can further eliminate the need for a trained expert to review.
- Improve the agreement rate (from the single-annotator approach) against the gold-standard approach from Cohen's κ of 0.68 to 0.77.

Table 1. Summary of annotation details described in a sample of existing HAR datasets where researchers mentioned using video or image annotation to obtain ground truth. Some researchers did not discuss components of their annotation approaches in detail (marked ✗), which include training annotators (Training), computing post-training inter-rater reliability (Reliability), and describing a post-annotation quality-control protocol (QC).

Dataset	Setting	Footage	Annotators	Training	Reliability	QC
CAPTURE-24 [4]	FL	2,562 hrs	A	✓	✗	✗
HARTH [9]	FL	37 hrs	A	✗	✓	✓
Opportunity [3]	Lab	5.3 hrs	A	✓	✗	✓
SPHERE [14]	Lab	✗	A	✗	✗	✗
Clemson [10]	Lab	15 hrs	A	✗	✗	✗
OxWalk [11]	FL	39 hrs	E	–	✗	✓
Hang-Time [6]	FL	38 hrs	A	✗	✗	✗

A: Trained annotators; E: Domain experts; FL: Free-living collection; Lab: Lab collection

– Result in more reliable annotations (compared to the commonly used single-annotator approach), improving Cohen's κ from 0.68 (for single-annotator annotations) to 0.79 (for silver-standard annotations).

2 Video annotation for human activity recognition

Researchers annotating datasets have often relied on a single-annotator approach [17]. We searched on Google Scholar using combinations of these keywords to identify publicly available HAR datasets: "datasets," "human activity recognition," and "wearable sensors." We reviewed the relevant references to identify related citations for publicly available datasets. Studies where the data are not available publicly for download were excluded. We identified only seven datasets, summarized in Table 1, where the researchers explicitly mentioned using video or image data to obtain ground-truth annotations. Ideally, reports on datasets would include descriptions that allow assessment of annotation quality, specifically information on training protocol, annotation protocol, inter-rater reliability, and quality-control procedures. In practice, HAR researchers often omit such information required to understand the provenance of the datasets they have released.

In the sample of relevant datasets we analyzed, only one dataset was annotated by trained domain experts and had an explicit quality control protocol; the two primary researchers in the OxWalk study [11] independently annotated the same 39 h of footage, but the paper does not include details on how disagreements were resolved. Other researchers have employed trained annotators, but researchers for only two such studies (CAPTURE-24 [4] and Opportunity [3]) mentioned their annotation training protocols. For example, the Opportunity study researchers organized sessions among annotators to discuss annotation edge cases in order to obtain reliable and consistent annotations. The CAPTURE-24 research team ensured that annotators were highly in agreement with an expert on example footage prior to annotating new data. Neither team reported post-training annotator reliability. Only one of the seven identified datasets includes data on post-training annotator reliability [9]. This sample of datasets suggests that even when collecting relatively small datasets, the gold-standard annotation protocol may not be used—likely due to the cost and burden involved—and details on how

Fig. 2. Comparisons between the single-annotator, silver-standard, and gold-standard approach (top to bottom). Annotation timelines were visualized using an hour of actual annotation data. Person-hours were calculated based on analyses using 16 h footage covering a waking day.

annotation was accomplished may not be available. Nevertheless, because annotated data are in short supply, such publicly available datasets are heavily used.

3 Method

Our own challenging annotation task led us to explore a middle-ground approach between the single-annotator approach and the gold-standard annotation approach, because gold-standard annotation is financially out of reach. Instead of having a second, completely independent annotator and a third domain expert resolving differences between annotators 1 and 2, we explored having a second trained annotator *review and revise* annotations from the first annotator (Figure 2). In this section, we describe the annotation training protocols and annotation method. Then we describe the experiments we used to assess the labor cost and annotation quality improvements of the proposed "silver-standard" annotation approach.

3.1 Annotation scheme and annotation software

Our taxonomy is designed to enable labeling of contextualized free-living activity and posture. Each annotation consists of at least two labels: the participant's posture (one of 13 mutually exclusive options, see Appendix Table 6) and physical activity type (one of 49 mutually exclusive options, see Appendix Table 7). Although we wanted our taxonomy to encompass both major HAR activities from prior research and fine-grained free-living behaviors, we balanced this comprehensiveness with practical constraints because annotators might have been overwhelmed by a taxonomy with a substantially larger set of labels. We opted for a taxonomy that would cover major activities from the American Time Use survey [15] but not specialized activities such as specific sports, activities that involve only one part of the body (e.g., hand gestures), or other highly specialized activities that were unlikely to occur among participants contributing to the dataset (e.g., sledding, jackhammering).

In our study, annotators labeled front-facing camera footage. The use of a front-facing camera might introduce additional ambiguity. One source of ambiguity results from the nature of the camera footage; distinguishing certain ambulation activities can be challenging, which contributed to a large portion of the unreliable annotations described later

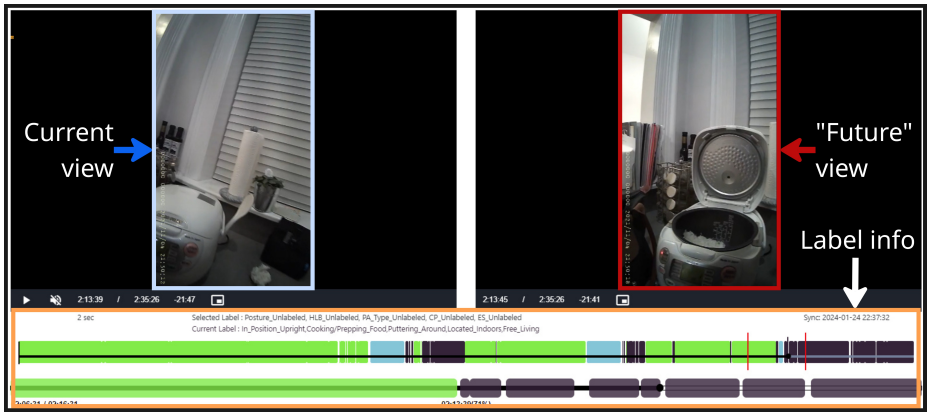


Fig. 3. A screenshot of the annotation software interface.

in Section 4.1. Another source of ambiguity is attributable to how participants wore the front-facing camera. If the cameras were partially covered, annotators sometimes cannot accurately decipher the current activity or posture. Thus, it is also important that the taxonomy allows annotators to indicate when labeling is not possible. In our annotation taxonomy, annotators can describe such cases with the use of *PA_ Type_ Indecipherable/ Video_ Unavailable* and *Posture_ Type_ Indecipherable/Video_ Unavailable*.

Researchers using our dataset will not be able to view the original front-facing camera video used for annotation because we must protect participants' privacy. To help researchers using the dataset interpret their results, our annotations also include non-mutually exclusive labels about the participant's "high-level behavior" (HLB) and additional context that may impact the participant's physical activity. The HLB and context labels are meant to provide helpful information but are not used for model training or evaluation directly, nor are they provided by any other datasets in Table 1, and thus we will not discuss quality assessment of these data in this work.

Our annotators use custom software designed for efficient free-living annotation using our taxonomy (Figure 3). Annotators can watch the video sped up when there are no changes in behavior, stopping at transitions to label. The "future" video view allows the annotator to anticipate changes in annotation complexity and dynamically change the video speed in response without overshooting, thus saving time. Annotators use the interface, with keyboard commands, to add new activity labels, merge two consecutive labels, split a label at the current frame, or edit the current labels. The software also flags behaviors that may be incompatible (e.g., a posture that is incompatible with an activity), provides suggestions of compatible posture or HLB labels (with the selected activity), and pre-populates selections with recent activities (Appendix Figure 4).

3.2 Annotation training protocol

We recruited annotators from the undergraduate student body of Northeastern University. Each completed a standardized training protocol. First, all annotators completed the required human subject research certification required by our institutional review board. Annotators then reviewed material on general data handling practices and the annotation process, including hands-on annotation practice under the supervision of a trained expert. After reviewing the material, they had to pass a quiz on the material.

Finally, to demonstrate annotation proficiency, annotators were required to achieve at least 90% second-by-second agreement in both the physical activity and posture label categories on a set of ten 10-minute videos annotated by a domain expert. Annotators were allowed to re-annotate these videos until they achieved the required benchmark. After reaching this benchmarking, annotators labeled two two-hour clips. The first consisted of a relatively common and obvious set of activities such as clear *Walking* or *Standing_With_Movement*; the second consisted of some less common activities such as *Doing_Resistance_Training_Free_Weights* and quick transitions between activities (e.g., short transitions between *Walking* and *Standing_With_Movement* while doing household activities such as cooking). Using the resulting annotations, we calculated post-training inter-rater reliability among all annotators. For the physical activity category, the newly trained annotators achieved an overall Fleiss' $\kappa = 0.76$ for the simple scenario and 0.61 for the complex scenario. While the resulting inter-rater reliability scores can both be interpreted as "substantial" agreement [8], the $\kappa = 0.61$ score for the complex scenario is closer to the "moderate" level agreement and demonstrates the complexity of our annotation task. The "moderate" level agreement on complex scenarios must be considered by future researchers as they develop HAR models using such data; ground-truth annotation will always contain some inaccurate or imprecise labels. These fast-changing scenarios, however, are not common in daily activities and should not hinder traditional HAR research aiming at recognizing more common daily ambulation activities, such as *Walking* or *Sitting_With_Movement*.

More intense or longer training might improve annotator reliability [1], but it also may be impractical. The entire annotation training protocol in our study, including the post-annotation reliability evaluation, requires about 20 h of effort for each annotator to complete. Further, our annotators are undergraduate students who are only available for a few months per semester and have limited working hours. The annotation task is also such a tedious task that many students do not want to continue annotating long-term. Thus, increasing annotation training time even more is not desirable, given that we do not anticipate large improvement gains.

3.3 Experiments

We explored three research questions (RQs) given our activity/posture taxonomy, front-facing video, and annotation software:

- **RQ1**: Can the silver-standard protocol produce annotations that are similar (i.e., high inter-rater agreement) to the gold-standard protocol?
- **RQ2**: Are the silver-standard annotations more reliable than single-annotator annotations?
- **RQ3**: Can annotators review and revise labels faster than annotating from scratch?

To explore these questions, we selected six videos covering six waking days totaling to 92 h of footage (from Monday until Saturday) from the first participant in our to-be-released dataset. During the entire waking day, this participant wore a front-facing camera (recording at 30 fps and 480p resolution) on their chest and only took the camera off for privacy reasons or during nighttime sleep.

Experiment 1: Silver-standard annotations versus gold-standard annotations To answer RQ 1 , we calculated the agreement rate between the silver-standard annotation and the gold-standard annotation. We selected one day with 16.4 h of footage; this is the day the participant performed the most activities.

First, to obtain the gold-standard annotation, independent annotators labeled each second of this video to obtain two independent annotation sets. Because the video is long and the annotation is laborious, multiple annotators were used for each set, but the result was two different independent annotations of the same day. A domain expert, also trained in annotation, then resolved disagreements from the two sets of annotations. The resolved annotations could contain activities that were not in either of the original annotation passes. The resulting annotation is considered the gold-standard annotation of the day. The domain expert required about 2 h of effort to resolve disagreements and obtain the final set of labels for 16.4 h of video.

To obtain the single-annotator annotation, a new set of independent annotators—*who were not involved in the process of generating the gold-standard annotation*—labeled the footage. We then obtained two sets of silver-standard annotations by selecting two trained annotators with varying experience levels annotating other data to be the "revisers." The less-experienced reviser had three months of experience, and the more experienced reviser had eight months of experience. All annotators work approximately 8–10 h per week for $15 USD per hour.

Experiment 2: Silver-standard annotations versus single-annotator annotations To answer RQ 2 and to assess whether the silver-standard approach is more reliable than the single-annotator approach, we evaluated the inter-rater reliability of silver-standard annotations compared to the single-annotator annotations on 92 h of data. Two independent annotator groups annotated all six days of videos resulting in two complete single-annotator annotation sets. We then recruited an additional group of revisers from our trained annotators–*who were not involved in the first pass of annotation*—to each individually revise these two sets of annotations. We then calculated inter-rater reliability between each of the single-annotator annotations as well as between each of the silver-standard annotations.

Analysis: Agreement metrics We considered each second of the video as a discrete segment for labeling. To evaluate inter-rater agreement, we used Cohen's κ as the primary metric and additionally reported percent agreement. The widely accepted interpretation of agreement ranges for Cohen's κ are <0.20 as "none to slight" agreement, 0.21—0.40 as "fair" agreement, 0.41—0.60 as "moderate" agreement, 0.61—0.80 as "substantial" agreement, and >0.8 as "almost perfect" agreement [8]. There is no equivalent interpretation for percent agreement. We used the Python implementation of Cohen's κ and percent agreement provided by scikit-learn version 1.6.

Analysis: Annotator productivity rate To answer RQ 3, we measured annotators' productivity rate while they were either annotating or revising. We define the *productivity rate* as the duration of footage annotated or revised divided by the total time spent working. Because annotators can work in multiple short shifts, we calculated and reported the mean productivity rate across all working shifts in that annotation set.

4 Results

For RQ 1 and RQ 2, we report inter-rater agreement on both the mutually exclusive physical activity and mutually exclusive posture types in Section 4.1 and 4.2, respectively. As mentioned previously, although annotators record additional high-level behavior and contextual information as they label, those labels are provided as part of the dataset

Table 2. Cohen's κ and percent agreement from the single-annotator and silver-standard annotations by less experienced (less-exp) and more experienced (more-exp) revisers against the gold-standard annotation in the physical activity category. To account for an unlikely error in the gold-standard annotation resulting from the two independent annotators both incorrectly labeling a segment of *PA_ Type_ Indecipherable/Video_ Unavailable* as *Sitting_ With_ Movement*, we calculated an adjusted κ and percent agreement while excluding this incorrect segment.

Annotation Set	Cohen's κ	Adjusted κ	% Agree	Adjusted %
Single-annotator annotation	0.68	0.71	80%	82%
Silver-standard (less-exp)	0.73	0.76	83%	85%
Silver-standard (more-exp)	0.77	0.81	86%	88%

only for additional understanding of the primary physical activity and posture labels and thus are not included in this analysis.

For RQ 3, we report the overall annotation and revision productivity rate in Section 4.3.

4.1 Silver-standard annotations versus gold-standard annotations

For the physical activity category, the single-annotator annotation was in "substantial" agreement with the gold-standard annotation with Cohen's $\kappa = 0.68$ (80% agreement). The silver-standard protocol further improved annotation quality, resulting in higher agreement rate with the gold-standard annotation. The more experienced reviser produced the better revised labels (i.e., higher agreement with gold-standard labels) with Cohen's $\kappa = 0.77$ (86% agreement), compared to our less experienced reviser with $\kappa = 0.73$ (82% agreement). While these agreement scores can all be interpreted as "substantial" agreement, the revised annotations were generally in better agreement with the gold-standard annotation.

One noticeable difference between the single-annotator, silver-standard, and gold-standard annotations was a long 1,600 s segment where the correct label was *PA_ Type_ Indecipherable/Video_ Unavailable*. In the gold-standard annotation, this portion was labeled as *Sitting_ With_ Movement*. The independent annotators likely inferred that the participant was *Sitting_ With_ Movement* because the participant was doing that activity immediately before and after the *PA_ Type_ Video_ Unavailable/Indecipherable* footage. Unfortunately, both our annotators introduced an incorrectly agreed label to the gold-standard annotation. Because we want to compare the silver-standard approach against the "best possible" annotation approach, we recomputed the "adjusted" Cohen's κ while excluding this 1,600 s segment from all annotation sets. The recomputed Cohen's κ was 0.71, 0.76, and 0.81 for single-annotator annotation, revision by less-experienced reviser, and revision by more-experienced reviser, respectively.

For the posture category, there were fewer available labels to choose from and there were less frequent posture changes. Labeling posture might be a simpler task. In fact, the single-annotator annotation and both silver-standard annotations by the two revisers were in "almost perfect" agreement with the gold-standard annotation (Appendix Table 8).

Changes made by the revisers Both revisers added details and new activities to the original annotation (Table 3). Both revisers added 16 min of the *Vacuuming* activity and 2 min of the *Washing_ Hands* activity. Both revisers also added more detailed bouts for ambulation activities that are frequently studied in HAR research. For example, the more experienced reviser added seven more bouts of *Walking_ Down_ Stairs*; even

Table 3. Changes made by the less experienced (Less E.) reviser and more experienced (More E.) reviser compared to the original (Orig.) annotation.

Activity	Total Duration (s)			Number of Bouts		
	Orig.	Less E.	More E.	Orig.	Less E.	More E.
Putting_ Clothes_ Away	0	218	0	0	2	0
Walking_ Down_ Stairs	82	82	107	11	13	18
Watering_ Plants	0	0	110	0	0	3
Loading_ Unloading_ Washer_ Dryer	0	0	122	0	0	3
Walking_ Up_ Stairs	138	140	149	15	15	17
Washing_ Hands	20	154	174	1	7	6
Kneeling_ With_ Movement	426	209	315	17	8	12
Sweeping	0	0	383	0	0	6
Folding_ Clothes	0	0	863	0	0	5
Vacuuming	0	982	967	0	1	1
Puttering_ Around	0	139	2,633	0	1	31
Walking	6,288	4,942	4,843	125	125	144
Standing_ With_ Movement	10,135	10,307	6,582	92	98	103
PA_ Type_ Unavailable	8,167	8,174	8,129	22	21	22
Sitting_ With_ Movement	33,324	33,349	33,323	25	25	27

though this was only 25 s of changed annotations. Similarly, this reviser added 19 more bouts of *Walking*, but reduced the total *Walking* duration by 1,445 s; in some health studies, measuring bouts of activity may be as important as volume of activity. The added bouts suggests that revisers were identifying, adding, or even refining the precision of quick ambulation activities.

The more experienced reviser added more new activities than the less experienced reviser, such as *Folding_ Clothes* (14.4 min), *Sweeping* (6.4 min), *Watering_ Plants* (2 min), and *Loading/Unloading_ Washing_ Machine/Dryer* (2 min). In total, the more experienced reviser added 43 min of new activities that were not included in the original annotation; the less experienced reviser added 22 min. The set of new activities added was different, with the less experienced reviser adding the *Putting_ Clothes_ Away* label but the more experienced reviser simply used the *Folding_ Clothes* label instead.

Changes not aligned with the gold-standard annotation Thirty-two percent of changes made by the reviser were not aligned with the adjusted gold-standard annotation. For example, the more experienced reviser made 1,903 s of changes that did not match (Appendix Table 10). Most of the differences (836 s) were due to the differences in the usage of two physical activity labels that represent different levels of motion: *Puttering_ Around* and *Standing_ With_ Movement*. In the taxonomy, *Puttering_ Around* is described as intermittently moving around and standing, while *Standing_ With_ Movement* is defined as being in a standing position where movements, such as shuffling feet or shifting body weight might occur. More precise annotators would label each individual movement bout instead of clustering them into *Puttering_ Around*. The experienced annotator was *more detailed* than the gold-standard annotation for a total of 196 s. Vice versa, the more experienced reviser used *Puttering_ Around* instead of using precise ambulation labels for 640 s. There were also 210 s of total non-aligned changes due to the reviser using precise activity labels (e.g., *Vacuuming* for 39 s) instead of generic ambulation labels. Finally, there were 60 s where the differences were from small boundary changes in activity start/stop times, or where the gold-standard annotation had a short gap in activity labels.

Table 4. Inter-rater reliability (Cohen's κ) computed using physical activity labels for single-annotator annotations and silver-standard annotations.

Day (video length)	**Single-annotator κ**	**Silver-standard κ**	**Improvement ($\Delta\kappa$)**
Day 1 (4.8 h)	0.49	0.77	+0.28
Day 2 (16.5 h)	0.83	0.83	0.00
Day 3 (17.7 h)	0.63	0.76	+0.13
Day 4 (19 h)	0.73	0.77	+0.04
Day 5 (17.4 h)	0.80	0.83	+0.03
Day 6 (16.4 h)	0.63	0.78	+0.15
Mean per day	**0.68**	**0.79**	**+0.11**

There were only 76 s of clear ambulation error between the silver-standard annotation (by the experienced reviser) and the gold-standard annotation. The experienced reviser also missed three activities: *Putting_ Clothes_ Away*, *Wet_ Mopping*, and *Organizing_ Shelf/Cabinet*, totaling 721 s. Notably, the reviser did change 5 min of *Standing_ With_ Movement* to *Folding_ Clothes*, but the gold-standard label was *Putting_ Clothes_ Away*.

4.2 Silver-standard annotations versus single-annotator annotations

On average across six days, the silver-standard physical activity annotations were more reliable than the single-annotator annotations with $\kappa = 0.79$ (in the "substantial" agreement range and close to being in "almost perfect" agreement) and $\kappa = 0.68$ ("substantial" agreement) respectively (Table 4). The biggest difference was for Day 1 with $\Delta\kappa = 0.28$. Across all days, the poorest agreement between independent silver-standard annotations was $\kappa = 0.76$ on Day 3, but was still an improvement from the single-annotator annotations on the same day with $\kappa = 0.63$.

Similarly, the silver-standard approach had consistent labeling for the posture category with mean (across six days) $\kappa = 0.91$ ("almost perfect" agreement); the single-annotator approach was in the "substantial" agreement range and came close to being "almost perfect," with the mean (across six days) $\kappa = 0.79$ (Appendix Table 9). Notably, single-annotator reliability for Day 1 was only "moderate" with $\kappa = 0.56$. All silver-standard posture annotations were in the "almost perfect" agreement range, with the minimum agreement at $\kappa = 0.88$. Given the higher agreement score across the board in both the physical activity and posture categories, we concluded that the silver-standard approach resulted in better annotation consistency over the single-annotator approach.

Total duration of all changes The revisers made between 1,866–10,822 s of total changes per day in the physical activity category (Appendix Table 11). For certain days, revisers changed up to 10,822 s, or more than 3 h, worth of changes. These changes made by revisers improved annotation quality by adding some new activities and more detailed activity bouts to the original annotation (Appendix Tables 12 and 13).

4.3 Productivity rate

Annotators revised annotations faster than they annotated from scratch (Table 5). Across our experiments, the average annotation rate was 2.2 h of annotation per hour worked (1 h of work resulted in 2.2 h of annotated video). Based on labeling efforts to date, including annotators who were not included in our experiments, the overall annotation rate is 2.8. In comparison, the mean revising rate was 5.6.

Table 5. Annotation rate (AR) and revision rate (RR), hours annotated/revised per hour worked, for annotators (of varying experience levels measured in months) in our experiments.

Annotator (experience)	**AR**	**RR**
Annotator 1 (3 mon)	1.1	3.5
Annotator 2 (8 mon)	3.1	7.6
Annotator 3 (3 mon)	1.5	4.4
Annotator 4 (3 mon)	2.1	5.9
Annotator 5 (3 mon)	2.6	5.9

Annotator (experience)	**AR**	**RR**
Annotator 6 (3 mon)	2.0	7.5
Annotator 7 (8 mon)	2.9	4.4
All annotators (N/A)	2.8	N/A
Mean (4.4 mon)	2.2	5.6

Based on the computed productivity rate, we estimated that, on average, it would take one annotator 7.3 h to annotate a typical video of a waking day (16 h). To achieve the gold-standard annotation, 15.6 h are required—14.6 h for two annotators to individually annotate, and 1–2 h for the domain expert to resolve disagreements (assuming 20% disagreement and a productivity rate of 2.8). The proposed silver-standard protocol would cut that time to 10.2 h, saving about 33% of total person-hours compared with the gold-standard approach.

5 Discussion

Demand for high-quality labeled data with increasingly large taxonomies that can be used to support machine learning and AI model training and validation is growing. As a result, a compromise between single-annotator annotation and the gold-standard method may be warranted. We are currently annotating a wearable sensor dataset with 22,000 h of front-facing video of free-living activity. Based on labeling to date, our trained annotators annotate at a mean annotation rate of 2.8 h of video annotated for every 1 h of annotation work. At this rate, we estimate that two independent annotation passes would take 15,800 person-hours. Additionally, a domain expert would need to resolve disagreements. Given a 20% disagreement rate (with $\kappa = 0.68$), this corresponds to 4,400 h of footage that the domain expert would need to resolve, which would take 1,600 h minimum. Overall, we estimate the gold-standard approach would require 17,400 person-hours of effort. In our study, the single-annotator inter-rater reliability can be as low as $\kappa = 0.61$ on more complex scenarios, which led us to explore the "silver-standard" method proposed here.

The complexity of our taxonomy contributes to the annotation challenge. In fact, our annotators were more accurate labeling postures because the annotation taxonomy only involves 13 postures, some of which are clearly distinguishable from the front-facing video. Furthermore, postures usually do not change as frequently as physical activities. In one minute, a person can be *Standing_With_Movement*, then *Sweeping*, then *Walking*, but still remain in the posture *In_Position_Upright* throughout. If our physical activity taxonomy were simplified (e.g., using a generic label to describe household ambulations), then annotation reliability might increase; yet, to improve HAR algorithms, the community needs more detailed labels for activities.

The nature of front-facing recording also contributes to the overall annotation challenge. Consider the *Puttering_Around* label, which was created to account for short, quick, and frequent changes between *Standing_With_Movement* and *Walking* for which precise annotation is not possible in a reasonable amount of time. In our training, we discouraged the use of *Puttering_Around*, but there are edge cases where participants moved their feet and shifted the weight of their body such that annotators felt the activity was in a gray area (such as when a person is moving around in a small kitchen for cooking). Also, sometimes it is not clear from the front-facing camera the degree

to which feet are moving. In our result described in Section 4.1, there were some differences in how this *Puttering_Around* label was used for labeling these edge cases; these differences likely contributed to the lower inter-rater agreement.

These unique annotation challenges led us to explore a compromised annotation approach between the single-annotator and the gold-standard approach. We decided to explore the silver-standard annotation strategy because the revision task might be inherently easier than the annotation task. When annotating from scratch, annotators have to look at both the "current" and "future" views (Figure 3) to identify when the current activity stops and a new activity starts. Juggling both views at the same time might require additional mental strain, on top of normal operations to add a new label. Revisers only need to look at the "current" view and determine if the provided label matches the current footage. Revising is, therefore, likely a simpler task, which is why revisers can pay more attention to finer details. In our experiments, the revisers improved the annotation by adding novel activities and more precise details that were missed using the single-annotator approach. These additional details and activities were small (in duration compared to the entire day), but for some studies might be important. Conventional ambulation activities, such as *Walking* or *Standing*, are already well-studied using lab data. Thus, the new frontier of activity recognition should be on detecting activities in free-living and detecting finer activities, including precisely detecting ambulation bouts. Currently, HAR algorithms are rarely evaluated on finer activities (e.g., *Putting_Clothes_Away*, or short bouts of *Walking_Up_Stairs*); high-quality annotated data are required to train and evaluate novel HAR algorithms that can precisely detect fleeting or uncommon activities in free-living individuals.

5.1 Limitations and implications

We conducted the experiments testing the silver-standard method on a week of front-facing video from a single participant (92 h). Our experiments required a combined total of 200 h of footage to be annotated and required 131 h of human labor. Although expensive, ideally, our experiments should be repeated on additional participants to further validate our results. Our investigation could further be enhanced with the collection and analysis of qualitative data from annotators explaining their thought processes as they revise annotations (e.g., were the revisers primarily fixing inaccurate annotations, or were they adding finer details that were originally missed? What strategies do they develop as they work?). Furthermore, our experiments did not evaluate how changes to the annotation taxonomy (e.g., larger taxonomies or taxonomies that include objects or hand gestures) might impact the utility of our proposed approach. The proposed revision strategy appears promising, but more work with a larger video sample set is warranted.

As one might expect, we found that more experienced annotators were more likely to produce better revisions. Unlike the gold-standard annotation approach, where the initial two annotations are independent, the revisers can be biased toward the original annotation when revising, thus producing poor-quality revisions. These poor-quality revisions, although unlikely to contain inaccurate annotations (i.e., using the wrong label to describe physical activity or posture), can fail to include finer activity details. More experienced annotators, who likely have a better understanding of the annotation scheme, might be more aware of fleeting activities. All revisions, however, were still better than the original single-annotator annotations. The revised annotations yielded much higher inter-rater agreement and were in high agreement with the gold-standard annotation.

Therefore, the silver-standard approach, although partially limited by the annotator's skill, is still likely to result in improved annotations over the original annotation.

One way to achieve better annotation might be longer, or more intensive, training [1, 18], but our training already involves 20 h of effort. In practice, most researchers delegate annotation to student annotators earning minimum wage [3, 18] who may have a high turnover rate that an onerous training process might further exacerbate. Another way to ensure better annotation quality might be to employ annotators long-term, because annotation quality might improve with practice [12], but the annotation task is tedious and not a task that many people want to continue long-term. A middle-ground approach may be to select a subset of high-performing annotators (e.g., annotators who performed well during the initial training) as revisers and to provide them with additional training specifically focused on optimizing annotation via revision.

Our annotators are paid $15/h USD, and a domain expert (doctoral student) is paid $40/h. Therefore, annotation of Day 6 of our test (16.4 h of footage) at actual annotation rates would run as follows: single-annotator annotation ($109.50 USD), silver-standard annotation ($153 USD), and gold-standard annotation ($299 USD). For our entire dataset, costs would be: single-annotator annotation ($118,500 USD), silver-standard annotation ($177,000 USD), and gold-standard annotation ($301,000 USD).

We have concluded that gold-standard annotation is out of reach for our project because of limited funding and resources. Had we sufficient resources, based on this work, we would use silver-standard annotation. Unfortunately, in practice, even achieving single-annotator annotation (using best practices for extensive annotator training) is proving challenging. We are now exploring additional strategies using computer-assisted annotation and revision.

6 Conclusion

We find that a "silver-standard" annotation strategy, using a single-pass annotation followed by a "revising" pass fix can yield annotation quality approaching the gold-standard level while saving 33% of total annotation time. The silver-standard annotation improved the agreement rate (compared to single-annotator annotation) against the gold-standard annotation from Cohen's κ of 0.77 up from 0.68. Furthermore, the silver-standard annotations had higher inter-rater reliability than single-annotator annotations, with Cohen's $\kappa = 0.79$ and 0.68, respectively. As demand for high-quality physical activity annotation data increases, researchers may want to consider strategies that augment single-annotator labeling and intensify efforts to report details of how annotation was accomplished and the effort that was required.

Acknowledgments Research reported in this publication was supported, in part, by the National Cancer Institute of the National Institutes of Health under award number R01CA252966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure The authors have no competing interests to declare that are relevant to the content of this article

References

1. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? A meta-analytic investigation. Computational Linguistics **37**(4), 699–725 (2011). https://doi.org/10.1162/COLI_a_00074

2. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychol. Sci. **6**(1), 3–5 (2011). https://doi.org/10.1037/14805-009
3. Calatroni, A., Roggen, D., Tröster, G.: Collection and curation of a large reference dataset for activity recognition. In: 2011 IEEE Int'l Conf. on Sys., Man, and Cybernetics. pp. 30–35. IEEE (2011). https://doi.org/10.1109/ICSMC.2011.6083638
4. Chan, S., Yuan, H., Tong, C., Acquah, A., Schonfeldt, A., Gershuny, J., Doherty, A.: CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. Scientific Data **11**(1), 1135 (2024). https://doi.org/10.1038/s41597-024-03960-3
5. Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M.H., White, T., van Hees, V.T., Trenell, M.I., Owen, C.G., Preece, S.J., Gillions, R., Sheard, S., Peakman, T., Brage, S., Wareham, N.J.: Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank study. PLoS ONE **12**(2), 1–14 (2017). https://doi.org/10.1371/journal.pone.0169649
6. Hoelzemann, A., Romero, J.L., Bock, M., Laerhoven, K.V., Lv, Q.: Hang-Time HAR: A Benchmark dataset for basketball activity recognition ising wrist-worn inertial sensors. Sensors (Basel) **23**(13) (2023). https://doi.org/10.3390/s23135879
7. Keadle, S.K., Lyden, K.A., Strath, S.J., Staudenmayer, J.W., Freedson, P.S.: A framework to evaluate devices that assess physical behavior. Exercise and Sport Sci. Reviews **47**(4), 206–214 (2019). https://doi.org/10.1249/JES.0000000000000206
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977). https://doi.org/10.2307/2529310
9. Logacjov, A., Bach, K., Kongsvold, A., Bårdstu, H.B., Mork, P.J.: HARTH: A human activity recognition dataset for machine learning. Sensors (Basel) **21**(23) (2021). https://doi.org/10.3390/s21237853
10. Mattfeld, R., Jesch, E., Hoover, A.: A new dataset for evaluating pedometer performance. In: 2017 IEEE Int'l Conf. on Bioinformatics and Biomedicine (BIBM). pp. 865–869 (2017). https://doi.org/10.1109/BIBM.2017.8217769
11. Small, S.R., Chan, S., Walmsley, R., von Fritsch, L., Acquah, A., Mertes, G., Feakins, B.G., Creagh, A., Strange, A., Matthews, C.E., et al.: Self-supervised machine learning to characterize step counts from wrist-worn accelerometers in the UK Biobank. Med. & Sci. in Sports & Exercise **56**(10), 1945 (2024). https://doi.org/10.1249/MSS.0000000000003478
12. Stoev, T., Suravee, S., Yordanova, K.: Variability of annotations over time: An experimental study in the dementia-related named entity recognition domain. In: INFORMATIK 2024. pp. 473–486. Gesellschaft für Informatik eV (2024). https://doi.org/10.18420/inf2024_35
13. Troiano, R.P., Berrigan, D., Dodd, K.W., Mâsse, L.C., Tilert, T., McDowell, M.: Physical activity in the United States measured by accelerometer. Med. & Sci. in Sports & Exercise **40**(1), 181 (2008). https://doi.org/10.1249/mss.0b013e31815a51b3
14. Twomey, N., Diethe, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P.: The SPHERE challenge: Activity recognition with multimodal sensor data. arXiv preprint (2016). https://doi.org/10.48550/arXiv.1603.00797
15. U.S. Bureau of Labor Statistics: American Time Use Survey. Data file (June 2024), https://www.bls.gov/tus/, accessed on July 28, 2025
16. Willetts, M., Hollowell, S., Aslett, L., Holmes, C., Doherty, A.: Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,609 UK Biobank participants. Scientific Reports **8**(7961) (2018). https://doi.org/10.1038/s41598-018-26174-1
17. Yordanova, K.: Challenges providing ground truth for pervasive healthcare system. IEEE Pervasive Computing **18**(2), 100–104 (2019). https://doi.org/10.1109/MPRV.2019.2912261
18. Yordanova, K., Kruger, F.: Creating and exploring semantic annotation for behaviour analysis. Sensors (Basel) **18**(9) (2018). https://doi.org/10.3390/s18092778

7 Appendix or Supplemental Material

Table 6. The posture taxonomy

Label	Description (The participant is definitely ...)
In_ Position_ Kneeling	In the position of kneeling, either still or with some other body movement. Could be with one knee or two.
In_ Position_ Reclining/Slouching	In the position of reclining or slouching back (typically on a chair, recliner, or couch). This would include reclining in bed if someone is propped up, but not *Lying_ On_ Back*.
In_ Position_ Sitting	In the position of sitting (includes sitting in a normal chair, barstool, swing, etc.).
In_ Position_ Upright	Upright, either still or moving. This posture includes both standing and ambulatory activities unless there is a distinct bending at the hip that substantially alters the upright position (e.g., bending to pick something up).
Lying_ On_ Back	In the position of lying on the back.
Lying_ On_ Left_ Side	In the position of lying on the left side.
Lying_ On_ Right_ Side	In the position of lying on the right side.
Lying_ On_ Stomach	In the position of lying on the stomach.
Posture_ Video_ Unavailable/ Indecipherable	The video is missing, too blurry, or at too poor of a camera angle to label the posture at this time.
Posture_ Other	In a specific and well-understood posture that would not be accurately described as being in a sitting, upright, lying (on back, left, right, stomach) or reclining/slouching posture. Examples might be well-defined Yoga or exercise postures (e.g., tree, hero, handstand) or on one's knees and doing something.
Posture_ Too_ Complex	The posture of the participant is deemed to be changing too fast or with too much complexity to label accurately during this time.
Posture_ Unlabeled	Posture for this segment of video has not been labeled. This is the default condition.

Table 7: The physical activity taxonomy

Label	Description (The participant is definitely ...)
Applying_ Makeup	Applying makeup to the face.
Bathing	Bathing in a tub. This does not include showering in a standing shower.
Blowdrying_ Hair	Using a hand-held blow dryer to dry hair. This does not involve hair drying at a parlor by a hairdresser.
Brushing_ Teeth	Cleaning teeth using a manual or electric brush.
Brushing/Combing/Tying_ Hair	Untangling/styling hair using a brush or a comb.
Cycling_ Active_ Pedaling_ Regular_ Bicycle	Pedaling while riding a non-stationary regular bicycle that facilitates the forward propulsion of the bicycle. This does not include coasting on the bicycle after pedaling or going downhill.
Cycling_ Active_ Pedaling_ Stationary_ Bike	Using cycle ergometer- e.g., commercially available exercise bikes such as a Peloton bike. Includes both seated and standing while pedaling. This does not include sitting on the bike at rest.

(continued on next page)

(continued from previous page)

Label	Description (The participant is definitely ...)
Doing_ Resistance_ Training_ Free_ Weights	Muscular training specifically involving the lifting of free weights. This label only applies to when the actual lifting (up/down) is occurring, not if the person is standing and resting.
Doing_ Resistance_ Training_ Other	All forms of resistance training other than lifting free weights. This includes using weight lifting machines and resistance bands or body weight.
Dry_ Mopping	Using a dry mop as an alternative to sweeping (e.g., Swiffer sweeper or other dry mop).
Dusting	Using a handheld duster to clean a surface. This ranges from using a cloth to other forms of modified dusters (e.g., a feather duster or Swiffer duster).
Flossing_ Teeth	Flossing teeth using string-based floss.
Folding_ Clothes	Folding clothes by hand. This includes laundered clothes and other cloth materials.
Ironing	Using a handheld iron or steamer to iron clothes.
Kneeling_ Still	In a kneeling position, still.
Kneeling_ With_ Movement	In a kneeling position but not still. may include upper or lower extremity movement while performing a task, fidgeting,or deliberate upper body movement (e.g., sway).
Loading/Unloading_ Washing_ Machine/Dryer	Manually placing/removing clothes from a washing machine.
Lying_ Still	In the lying posture with no limb or body movement.
Lying_ With_ Movement	In the lying posture with some limb or body movement. This may typically involve limb movement that occurs when doing an activity while lying, e.g, flipping pages when reading a book, or gesturing while speaking on the phone with someone.
Organizing_ Shelf/Cabinet	Arranging/rearranging items in a shelf/cabinet/bookcase to organize or tidy up.
PA_ Type_ Other	A specific, well-understood PA_Type that would not be accurately described as any of the others on this list.
PA_ Type_ Too_ Complex	The behavior of the participant is deemed to be changing too fast or with too much complexity to label accurately during this time.
PA_ Type_ Unlabeled	PA_Type for this segment of video has not been labeled. This is the default condition.
PA_ Type_ Video_ Unavailable/ Indecipherable	The video is missing, too blurry, or at too poor of a camera angle to label the posture at this time.
Playing_ Frisbee	Playing with a flying disc with someone else, or outside with a dog.
Puttering_ Around	Upright and intermittently moving around/standing, such as often happens when cooking. If in the middle of *Puttering_ Around* there is a clear, extended bout of another activity that should be explicitly labeled.
Putting_ Clothes_ Away	Stowing clothes in a closet, cabinet, or other storage area.
Running_ Non-Treadmill	Running outdoors or on an indoor track or in any other context that is not a treadmill.
Running_ Treadmill	Running on a treadmill.
Showering	Showering in a standing shower. This does not include bathing in a tub.
Shoveling_ Mud/Snow	Using a hand-held shovel to manually move snow or mud for any purpose.

(continued on next page)

(continued from previous page)

Label	Description (The participant is definitely ...)
Sitting_ Still	In a seated position, still. Includes sitting at rest on a stationary bicycle.
Sitting_ With_ Movement	In a seated position but not still. may include upper or lower extremity movement while performing a task, fidgeting,or deliberate upper body movement (e.g., sway). Includes coasting on a bicycle while seated.
Standing_ Still	In a standing position, still.
Standing_ With_ Movement	In a standing position but not still. Movement may include upper or lower extremity movement while performing a task, fidgeting, or deliberate upper body swaying. Normal shuffling of feet while standing in the same place (e.g., to shift weight or move forward slowly in a line) may occur.
Sweeping	Using a handheld broom to sweep the floor.
Synchronizing_ Sensors	Performing the sensor and camera time syncing activity specific to the Datasets Project. This involves clapping of hands and flexing the hip joint such that each knee is raised with the help of one's hands.
Vacuuming	Using an electric vacuum to clean a floor/carpet. This does not include small portable handheld vacuums.
Walking_ Down_ Stairs	Going downstairs. This may involve a break in descending that occurs when walking on a level surface for brief bouts between consecutive flights of stairs. It could involve the use of handrails for support.
Walking_ Fast	Walking faster than a normal pace that is deliberate. E.g., purposefully increasing walking speed to make sure one can catch a train, or to get somewhere on time. A normal walking speed is between 3 and 3.5 mph for most people. This is a judgment call and should be used only if you are convinced that this is deliberate.
Walking_ Slow	Walking slower than a normal pace that is deliberate. E.g., strolling in a park and chatting with someone. A normal walking speed is between 3 and 3.5 mph for most people. This is a judgment call and use only if you are convinced that this is deliberate. Use unless it is not discernibly uphill. If so, label as <Walking>.
Walking_ Treadmill	Walking on a treadmill.
Walking_ Up_ Stairs	Going up stairs. This may involve a break in ascending that occurs when walking on a level surface for brief bouts between consecutive flights of stairs. It could involve the use of handrails for support.
Walking	Walking naturally.
Washing_ Face	Washing one's face using water, soap, or face wash. This includes removing makeup.
Washing_ Hands	Deliberately washing one's hands using water, soap, or hand sanitizer.
Watering_ Plants	Watering household plants using a watering can or hose.
Wet_ Mopping	Using a handheld mop or cloth to clean the floor. Includes manually mopping the floor using a cloth, mop and bucket, electronic handheld mops such as Swiffer Wetjet. These are different from dry mopping as the latter involves more activities (e.g., dipping mop in bucket, wringing, mopping) and requires significantly higher effort.

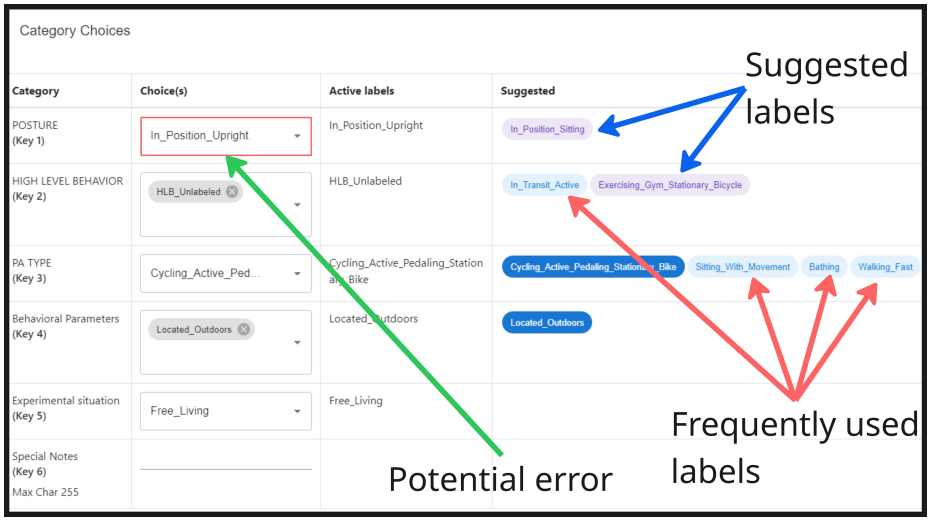


Fig. 4. A screenshot of the label edit dialogue.

Table 8. Cohen's κ and percent agreement from the single-annotator and silver-standard annotations by less experienced (less-exp) and more experienced (more-exp) revisers against the gold-standard annotation in the posture category. To account for an unlikely error in the gold-standard annotation resulting from the two independent annotators both incorrectly labeling a segment of *Posture_Indecipherable/Video_Unavailable* as *In_Position_Sitting*, we calculated an adjusted κ and percent agreement while excluding this incorrect segment.

Annotation Set	Cohen's κ	Adjusted κ	% Agree	Adjusted %
Single-annotator annotation	0.86	0.94	94%	96%
Silver-standard (less-exp)	0.90	0.95	94%	97%
Silver-standard (more-exp)	0.86	0.90	92%	94%

Table 9. Inter-rater reliability (Cohen's κ) computed using posture labels for single-annotator annotations and silver-standard annotations.

Day (video length)	Single-annotator κ	Silver-standard κ	Improvement
Day 1 (4.8 h)	0.56	0.88	+0.32
Day 2 (16.5 h)	0.94	0.95	+0.02
Day 3 (17.7 h)	0.70	0.90	+0.20
Day 4 (19 h)	0.93	0.95	+0.02
Day 5 (17.4 h)	0.84	0.88	+0.04
Day 6 (16.4 h)	0.78	0.89	+0.11
Mean per day	**0.79**	**0.91**	**+0.12**

Table 10: Details of all non-aligned changes made by the more experienced reviser discussed in Section 4.1

Changed From	Changed To	Should Be	Duration (s)
Reviser Less Detailed and Used Puttering_Around (Total: 640 s)			
Standing_With_Movement	Puttering_Around	Standing_With_Movement	458
Walking	Puttering_Around	Standing_With_Movement	161

Continued on next page

Table 10 – continued from previous page

Changed From	Changed To	Should Be	Duration (s)
Walking	Puttering_Around	Walking	15
PA_Type_Unlabeled	Puttering_Around	Standing_With_Movement	6
Actual Ambulation Disagreement (Total: 76 s)			
Standing_With_Movement	Walking	Standing_With_Movement	60
Walking	Standing_With_Movement	Walking	8
Standing_With_Movement	Walking	Sitting_With_Movement	3
Sitting_With_Movement	Walking	Sitting_With_Movement	3
PA_Type_Unlabeled	Standing_With_Movement	Walking	2
Annotator Did Not Use Puttering_Around and Was More Detailed (Total: 196 s)			
Standing_With_Movement	Walking	Puttering_Around	111
Walking	Standing_With_Movement	Puttering_Around	37
Sitting_With_Movement	Walking	Puttering_Around	14
PA_Type_Unlabeled	Walking	Puttering_Around	12
Kneeling_With_Movement	Walking	Puttering_Around	10
Walking_Up_Stairs	Walking	Puttering_Around	7
PA_Type_Unlabeled	Standing_With_Movement	Puttering_Around	5
Used Detail Activity Instead of Ambulation Related (Total: 210 s)			
Standing_With_Movement	Sweeping	Puttering_Around	37
Walking	Vacuuming	Standing_With_Movement	29
Standing_With_Movement	Washing_Hands	Standing_With_Movement	26
Standing_With_Movement	Sweeping	Standing_With_Movement	23
Standing_With_Movement	Washing_Hands	Puttering_Around	16
Walking	Walking_Down_Stairs	Walking	12
Walking	Watering_Plants	Standing_With_Movement	12
Standing_With_Movement	Watering_Plants	Puttering_Around	8
Walking	Vacuuming	Puttering_Around	7
Walking	Watering_Plants	Puttering_Around	7
Standing_With_Movement	Loading/Unloading_Washer_Dryer	Standing_With_Movement	7
Standing_With_Movement	Folding_Clothes	Walking	7
Kneeling_With_Movement	Sweeping	Puttering_Around	5
Walking	Walking_Down_Stairs	Puttering_Around	4
Standing_With_Movement	Folding_Clothes	Standing_With_Movement	3
Walking	Kneeling_With_Movement	Puttering_Around	3
PA_Type_Unlabeled	Vacuuming	Puttering_Around	2
Kneeling_With_Movement	Loading/Unloading_Washer_Dryer	Standing_With_Movement	2
Annotator Missed a Detailed Activity (Total: 721 s)			
Standing_With_Movement	Folding_Clothes	Putting_Clothes_Away	310
Standing_With_Movement	Puttering_Around	Organizing_Shelf/Cabinet	159
Standing_With_Movement	Walking	Putting_Clothes_Away	92
Walking	Puttering_Around	Putting_Clothes_Away	82
Standing_With_Movement	Puttering_Around	Putting_Clothes_Away	61
Standing_With_Movement	Puttering_Around	Washing_Hands	8
Walking	Puttering_Around	Folding_Clothes	7
PA_Type_Unlabeled	Walking	Putting_Clothes_Away	2
Missing Label or Boundary Related Error (Total: 60 s)			
Standing_With_Movement	Puttering_Around	PA_Type_Unlabeled	9
Standing_With_Movement	Walking	PA_Type_Unlabeled	8
Walking	Puttering_Around	PA_Type_Video_Unavailable	8
Standing_With_Movement	Folding_Clothes	PA_Type_Video_Unavailable	7
Sitting_With_Movement	Walking	PA_Type_Unlabeled	5
PA_Type_Unlabeled	Walking	PA_Type_Unlabeled	3
PA_Type_Unlabeled	Puttering_Around	PA_Type_Unlabeled	3
Standing_With_Movement	Sweeping	PA_Type_Unlabeled	2
Kneeling_With_Movement	Loading/Unloading_Washer_Dryer	PA_Type_Video_Unavailable	2
Standing_With_Movement	Walking	PA_Type_Video_Unavailable	2
Standing_With_Movement	Walking	Kneeling_With_Movement	1
PA_Type_Unlabeled	Walking	Walking_Up_Stairs	1
Walking_Up_Stairs	Walking	Walking_Up_Stairs	1
Standing_With_Movement	Vacuuming	Standing_With_Movement	1
Walking	PA_Type_Unlabeled	Sitting_With_Movement	1
Standing_With_Movement	Puttering_Around	Walking	1
Standing_With_Movement	Washing_Hands	PA_Type_Unlabeled	1
PA_Type_Unlabeled	Sitting_With_Movement	Sitting_Still	1
PA_Type_Unlabeled	Sitting_With_Movement	Walking	1
Walking_Down_Stairs	Walking_Up_Stairs	Walking	1
Standing_With_Movement	PA_Type_Unlabeled	Puttering_Around	1
		Total Duration of All Incorrect Fixes	**1,903 s**

Table 11. Duration of changes made by revisers (in seconds) across the two independent annotator groups.

Day (video length)	First annotator group (s)	Second annotator group (s)
Day 1 (4.8 h)	254	4,417
Day 2 (16.5 h)	1,866	3,285
Day 3 (17.7 h)	6,784	4,254
Day 4 (19 h)	2,130	7,687
Day 5 (17.4 h)	4,839	10,822
Day 6 (16.4 h)	7,621	10,355

Table 12. Changes made by the reviser compared to the original annotation by the first annotator group.

Activity	Total Duration (s)		Number of Bouts	
	Original	Revised	Original	Revised
Walking_Fast	20	0	1	0
Walking_Slow	10	10	1	1
Standing_Still	73	51	2	1
PA_Type_Too_Complex	0	55	0	1
PA_Type_Other	116	122	8	6
Watering_Plants	242	230	1	1
Brushing_Teeth	251	242	3	3
Synchronizing_Sensors	236	264	9	10
Walking_Down_Stairs	187	301	25	41
Loading/Unloading_Washing_Machine/Dryer	334	334	4	4
Walking_Up_Stairs	208	391	29	47
Lying_Still	411	411	1	1
Ironing	436	436	1	1
Sweeping	447	507	2	3
Organizing_Shelf/Cabinet	549	549	2	2
Putting_Clothes_Away	667	644	6	5
Folding_Clothes	657	689	7	7
Kneeling_With_Movement	667	696	14	16
Washing_Hands	727	823	25	33
PA_Type_Unlabeled	844	844	4	4
Vacuuming	1,035	1,035	1	1
Cycling_Active_Pedaling_Stationary_Bike	9,448	9,448	2	2
Walking	13,745	13,143	214	242
Puttering_Around	26,860	22,252	319	487
Standing_With_Movement	16,271	22,288	267	444
PA_Type_Video_Unavailable/Indecipherable	52,854	62,704	71	101
Sitting_With_Movement	199,094	190,279	229	229

Table 13. Changes made by the reviser compared to the original annotation by the second annotator group.

Activity	Total Duration (s)		Number of Bouts	
	Original	Revised	Original	Revised
Cycling_Active_Pedaling_Regular_Bicycle	0	27	0	1
Standing_Still	31	31	1	1
Synchronizing_Sensors	70	70	2	2
Folding_Clothes	0	156	0	1
PA_Type_Other	2,404	172	49	13
Loading/Unloading_Washing_Machine/Dryer	0	187	0	5
Brushing_Teeth	261	261	3	3
Putting_Clothes_Away	347	355	3	3
Walking_Down_Stairs	176	381	17	33
Walking_Up_Stairs	304	553	27	42
Washing_Hands	462	606	18	24
Kneeling_With_Movement	416	710	7	16
Vacuuming	0	971	0	1
Sitting_Still	989	989	5	5
Puttering_Around	24,805	4,660	372	111
Cycling_Active_Pedaling_Stationary_Bike	5,700	9,289	1	4
Walking	19,380	20,684	200	413
Standing_With_Movement	25,772	39,093	414	508
PA_Type_Video_Unavailable/Indecipherable	60,118	61,315	120	132
Sitting_With_Movement	188,371	189,223	444	441