“Fine-grained Activities of People Worldwide”
Supplementary material, WACV 2023

Appendix

A. Collector Platform

Collector is a global platform for collecting large scale consented video datasets of people for visual AI applications. Collector is able to record, annotate and verify custom video datasets of rarely occurring activities for training visual AI systems. The Collector platform provides:

- On-demand collection of rarely occurring activities.
- Simultaneous video recording, annotation and verification into a single unified platform.
- Touchscreen UI for live annotation of bounding boxes, activity clips and object categories.
- Specification of required collection attributes such as pose, illumination, location or object interactions.
- IRB approved informed consent for ethical dataset construction with in-app face anonymization.

In this section, we will describe the motivation and design of the Collector platform. This platform includes a *mobile app* for collecting labeled videos sourced from freelancers around the world, a *campaign dashboard* for setup, control and monitoring of a world-wide collection campaign by a dataset administrator and *quality control* for human review and distributed consensus for annotation quality. This platform was used for collection of the dataset in section 4, and we believe will be useful for the research community to support future dataset collection.

A.1. Mobile App for Recording and Annotation

The front end of the Collector platform is a free mobile app designed to streamline consenting, recording, annotation and verification from collectors around the world.

Figure 3 shows an overview of the collector workflow. Collectors are invited onto the platform, and they download the collector mobile app to their device. Collectors are presented *collections* which are video collection tasks grouped by required objects (e.g. a car, a motorcycle) or locations (e.g. parking lot, dining room). Each collection specifies the requirements of the submitted video, which include required activities, objects, location, illumination conditions, actor pose and camera viewpoint. Once a collector chooses a collection to record, they get consent from their subject, including a video recording to ensure that the person consenting is the person being recorded. Collector recruitment requires proficient readers of English in order to provide us informed consent in this step. Next, the collector watches an example video which shows a gold standard exemplar of the collection. We use visual exemplars to bypass language issues and communicate an idea of what the collection should look like. Finally, the collector records and annotates the video live using touch gestures on their device, corrects errors using an in-app annotation editor and submits the annotated collection for review. Annotations include bounding boxes around objects, object labels and start and end times for each activity in the collection, all collected while the video is being recorded. The best submissions from our worldwide collection team as adjudicated by the review team are used as new training examples for newer collectors. In other words, collectors “see one, do one, and teach one” on our platform.

Figure A.11 shows an example of the in-app annotation editor. This editor is used to annotate videos in-app after they have been collected. This is useful for collecting data to support the Activity Detection (Rigid) task, where the device must be stationary during recording, where collector cannot annotate in-app while they record. Annotations include bounding boxes for objects and people, which is specified using multi-touch gestures for fast video annotation. Further annotations include start and end times for activities, which are specified by press gestures in a bounding box for the start (press-down) and end (lift) when an activity occurs. The saved edited video is uploaded to the Collector backend for further processing.

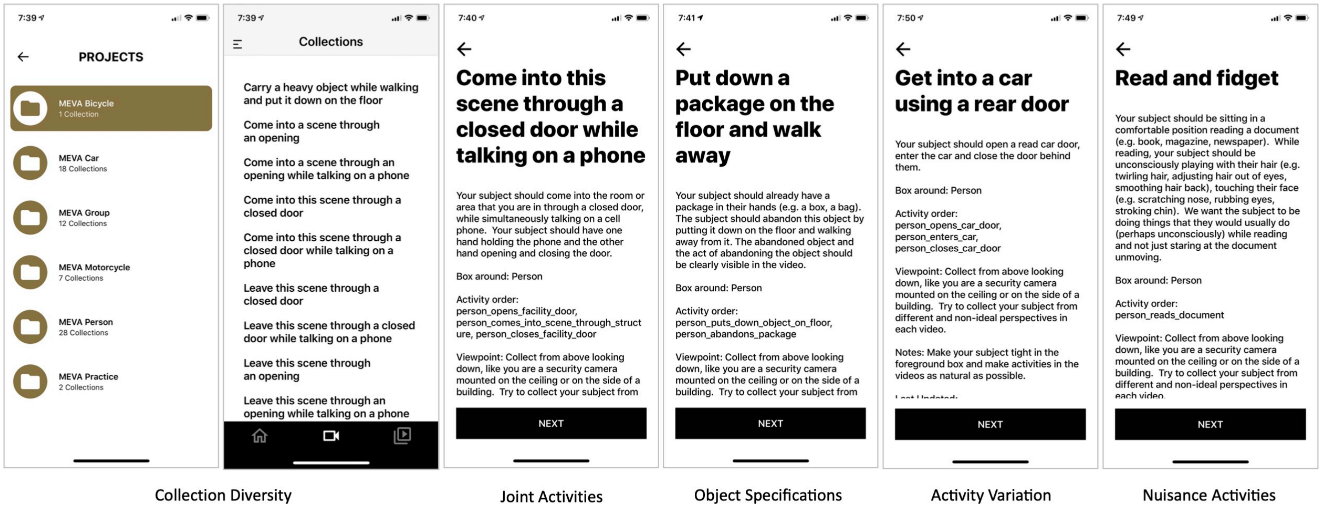
Figure A.10 shows an example of increasing the diversity of collections by controlling the collections available in the campaign. These screenshots show what is presented to the collectors in-app when they are tasked with collecting diverse data. The key component for this workflow is showing the collectors an example video for what should attempt to collect along with a written description. This provides multiple resources to the collectors to aid them in collecting high quality data of the form needed by the campaign.

The Collector mobile app is freely available in the iOS and Android app stores. This app has been downloaded and used by thousands of freelance collectors worldwide. More information and a tutorial video for collector usage is available at visym.com/collector.

A.2. Campaign Dashboard for Global Coordination

Large scale dataset collection includes a large volume of videos to be collected, annotated and verified. Each collection campaign may contain hundreds of different collection types, and each collection type may have tens of thousands of video submissions from all around the world. Campaign management tools are critical to coordinate the workforce, monitor submissions for fraud or unclear instructions, and maintain high quality for this volume of data.

The collector platform coordinates a global workforce using a *campaign dashboard*. The campaign dashboard provides a real-time interface for an administrator to set up,



Carry a heavy object while walking and put it down on the floor

Come into a scene through an opening

Come into a scene through an opening while talking on a phone

Come into this scene through a closed door

Come into this scene through a closed door while talking on a phone

Leave this scene through a closed door

Leave this scene through a closed door while talking on a phone

Leave this scene through an opening

Leave this scene through an opening while talking on a phone

Collection Diversity

Come into this scene through a closed door while talking on a phone

Your subject should come into the room or area that you are in through a closed door, while simultaneously talking on a cell phone. Your subject should have one hand holding the phone and the other hand opening and closing the door.

Box around: Person

Activity order:
person_opens_facility_door,
person_comes_into_scene_through_structure, person_closes_facility_door

Viewpoint: Collect from above looking down, like you are a security camera mounted on the ceiling or on the side of a building. Try to collect your subject from

Joint Activities

Put down a package on the floor and walk away

Your subject should already have a package in their hands (e.g. a box, a bag). The subject should abandon this object by putting it down on the floor and walking away from it. The abandoned object and the act of abandoning the object should be clearly visible in the video.

Box around: Person

Activity order:
person_puts_down_object_on_floor,
person_abandons_package

Viewpoint: Collect from above looking down, like you are a security camera mounted on the ceiling or on the side of a building. Try to collect your subject from

Object Specifications

Get into a car using a rear door

Your subject should open a read car door, enter the car and close the door behind them.

Box around: Person

Activity order:
person_opens_car_door,
person_enters_car,
person_closes_car_door

Viewpoint: Collect from above looking down, like you are a security camera mounted on the ceiling or on the side of a building. Try to collect your subject from different and non-ideal perspectives in each video.

Notes: Make your subject tight in the foreground box and make activities in the videos as natural as possible.

Last Updated:

Activity Variation

Read and fidget

Your subject should be sitting in a comfortable position reading a document (e.g. book, magazine, newspaper). While reading, your subject should be unconsciously playing with their hair (e.g. twirling hair, adjusting hair out of eyes, smoothing hair back), touching their face (e.g. scratching nose, rubbing eyes, stroking chin). We want the subject to be doing things that they would usually do (perhaps unconsciously) while reading and not just staring at the document unmoving.

Box around: Person

Activity order:
person_reads_document

Viewpoint: Collect from above looking down, like you are a security camera mounted on the ceiling or on the side of a building. Try to collect your subject from different and non-ideal perspectives in each video.

Nuisance Activities

Figure A.10. Collection diversity on the Collector platform. The platform allows for controlling the diversity of collections, by introducing collection variants with simultaneously occurring activities, new objects, within-class variation or rarely occurring activities.

control and monitor a campaign. The dashboard enables an administrator to:

- **Plan**. The admin defines the campaign by specifying each collection, along with the required objects, viewpoints, styles, illuminations, locations and IRB approved consent language.
- **Train**. The admin team collects one or more reference videos for each collection. Each collector is shown the description of a collection, and they view one or more reference videos for this collection. This provides a “see one, do one” strategy for non-native English speakers.
- **Deploy**. The campaign is deployed based on a collection schedule, including a maximum number of collections allowed per collector and per campaign. Admins can onboard and offboard collectors worldwide.
- **Monitor**. The dashboard shows a view live submission stream from collectors and reviewers worldwide. This allows the dataset admin to confirm that the collection is being submitted correctly and successfully.
- **Refine**. The dataset admin can refine the collection or training videos based on the performance of the the collector community. For example, appendix figure A.10 shows examples of refining a collection to increase diversity of objects and activities.
- **Pay**. Freelance collectors worldwide are paid per video that is submitted and accepted by the review team. Payments are made on a bi-weekly basis.

Figure A.12 shows an example of the live dashboarding providing the state of the campaign. This dashboard is the primary visualization of the state of the collection campaign that is used by an administrator for monitoring, command and control. This captures a global near real-time view of the worldwide submissions along with the ability to visualize submissions from any collector or reviewer.

A.3. Human Review for Annotation Quality

Large scale dataset collection requires careful review of videos to maintain dataset quality. In the Collector platform, the review team is tasked with daily verification of submissions to check that they satisfy the collection requirements. Reviewers are promoted from within the worldwide collector pool, and are selected based on their historical submission quality, and their interest in reviewing the work of fellow collectors in exchange for a fixed price per review.

Figure A.13 shows the reviewing interface. These screen shots show what is presented with reviewers when they are tasked with maintaining quality control. We have developed an HTML based interface for quickly reviewing the annotation quality of a video submitted by the collectors. Reviewers are sent a daily email with an HTML review link containing their reviews for the day. This review interface displays animated WEBP clips, in a montage format that allows for fast reviewing of a video at a glance. Reviewers are tasked with (i) reading the collection description, (ii) watching the reference example video for this collection, (iii) watching the submitted video and (iv) selecting one or more reviews by pressing the appropriate HTML button.

Each video is reviewed by a fixed number of reviewers (usually three), and a mean quality score is assigned to each video. Reviews are streamed to the Collector backend for aggregation, and reviewers are paid a fixed fee for each review submitted. Reviewers are audited by providing them reviews in their review stream with a known label (e.g. a synthetically corrupted video or a video reviewed by an administrator), to check their review quality and give them feedback. Finally, if this video quality is above a campaign specified threshold, then it is authorized for payment.

B. Consented Activities of People Dataset

In this section, we discuss the related work, key challenges, collection methodology and distribution format for curating a large scale dataset of activities of daily life. Section B.3 specifies the design challenges of this dataset and motivates our design goals. Section B.4 describes the collection parameters using the Collector platform. Finally, section B.5 describes the format of the dataset and the evaluation tasks.

B.1. Related Work

Table 1 shows a dataset comparison with the state of the art. For datasets with multiple evaluation tasks, we select the task and associated data most closely related to activity classification or activity detection. For example, Ego4D has five benchmark tasks, and we compare with the subset of data labeled for the Moment Query (MQ) task. For those datasets with label space organized as a multi-level hierarchy (e.g. ActivityNet) we select the lowest level as the number of fine classes and the immediate parents as the number of coarse classes. If there is no hierarchical organization, we report the number of classes as coarse classes. Clips reports the number of instances of each class (e.g. a trimmed clip containing an activity) available for testing, validation or training. We show only those mean clips per class that are reported by the authors in the source publication. The CAP dataset reports the mean clips per class across all classes, and mean clips per class considering only the top-250 fine-classes with the largest number of instances, as shown in figure 4 (left). Finally, note that this analysis does not include specialized domains such as fine-grained activities in sports datasets [17][40][54][50][63][68][53][49].

This table shows that the proposed CAP dataset is the largest consented dataset of people as measured by mean clips per class for training.

B.2. Design Objectives

The CAP dataset has the following design objectives:

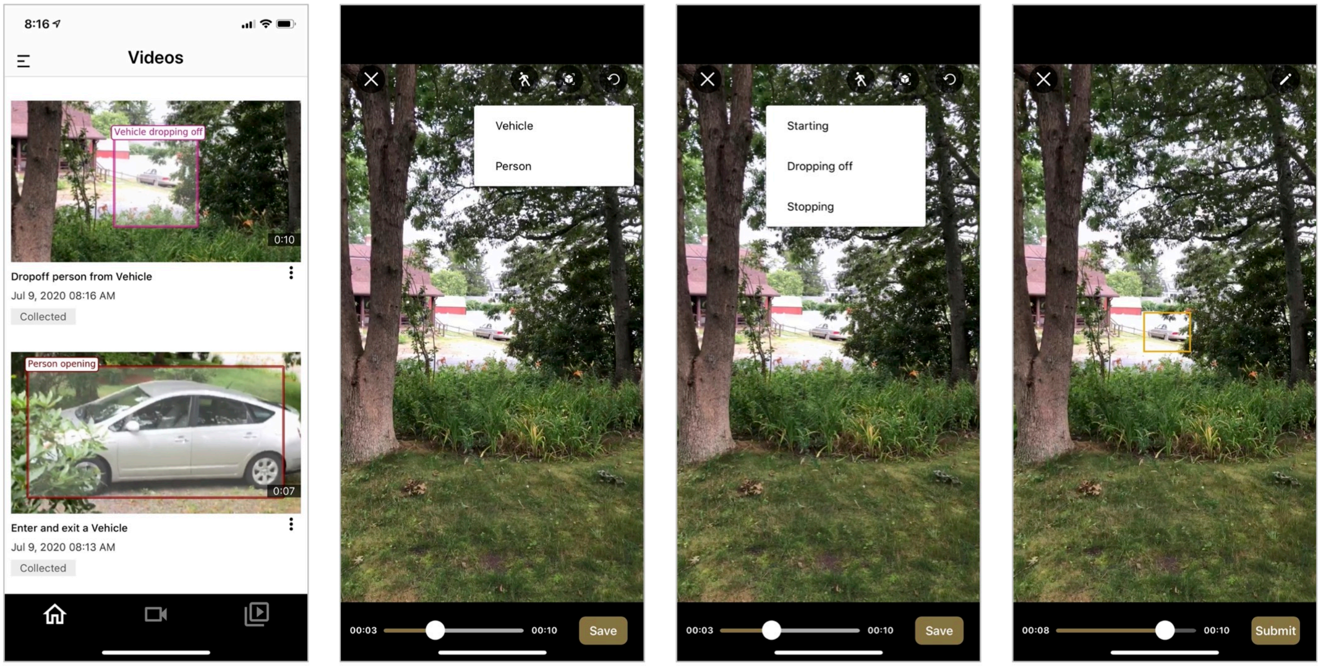
- **Atomic.** Activities should be short duration with length $\leq$ 3 seconds and visually grounded (e.g. activities should be discriminative from the pixels).
- **Fine-grained.** Activities should be selected where motion is critical for discrimination, rather than the scene context or object appearance.
- **Daily-life.** The collection should involve locations, activities and objects that people use or perform every day, without practice or expertise.
- **Non-overlapping.** All activities should be performed independently. (e.g. a subject will not simultaneously use a cell phone while taking off a hat).
- **Person-centered**. All videos should include a primary consented person performing the activity.
- **Third-person**. All videos should be collected from a third-person viewpoint, looking down on the scene from above, consistent with a ceiling/wall mounted camera.
- **Diverse**. Activities should be collected to encourage diversity in culture, geographic location, viewpoint, objects, pose and illumination.
- **Worldwide.** Videos should be collected from many countries around the world.
- **Ethical.** All videos must be collected with informed consent for how the videos will be shared and used.
- **Balanced**. Activities should be collected so that the number of instances per class is approximately equal, including labels that are rare in natural video.
- **Large-scale.** The dataset should include a liberal license with open distribution format and easily downloadable training and validation set.
- **Annotated**. Videos should be annotated with bounding box tracks around the primary actor along with temporal start/end frames for each activity instance.

B.3. Design Challenges

These design objectives introduce a number of challenges and open questions. What is a fine-grained activity? How are the activity labels selected? How can we collect balanced data of infrequent labels? How do we control the diversity of the data collected? How do we ensure a dataset is collected both globally and ethically?

What are fine-grained activities? Fine-grained visual categorization is an established task in the object recognition literature [23, 15, 75, 66, 58, 77, 52, 44, 12]. The term “fine-grained label” or “fine grained category” was originally introduced in the context of image classification of subordinate object categories, such as bird species, plant species or product brands. These are classes with subtle discriminative features, such as the color of a wingtip or shape of a leaf. These annotations often require an expert to specify the class label, and the differences between classes are subtle, highly localized and require expert training. However, the differences between classes are visually grounded in the pixels, if you know where to look.

A fine-grained activity is not as straightforward to define. An activity is typically defined by a verb being performed by a noun. For example, the activity *person sits* has the noun “person” performing the verb “sits”. Is this a fine-grained category? Compare the activity category of *person sits* to *person squats* vs. *person drinks*. In the first case, there are



Collected Videos

Object Editor

Activity Editor

Editor Preview

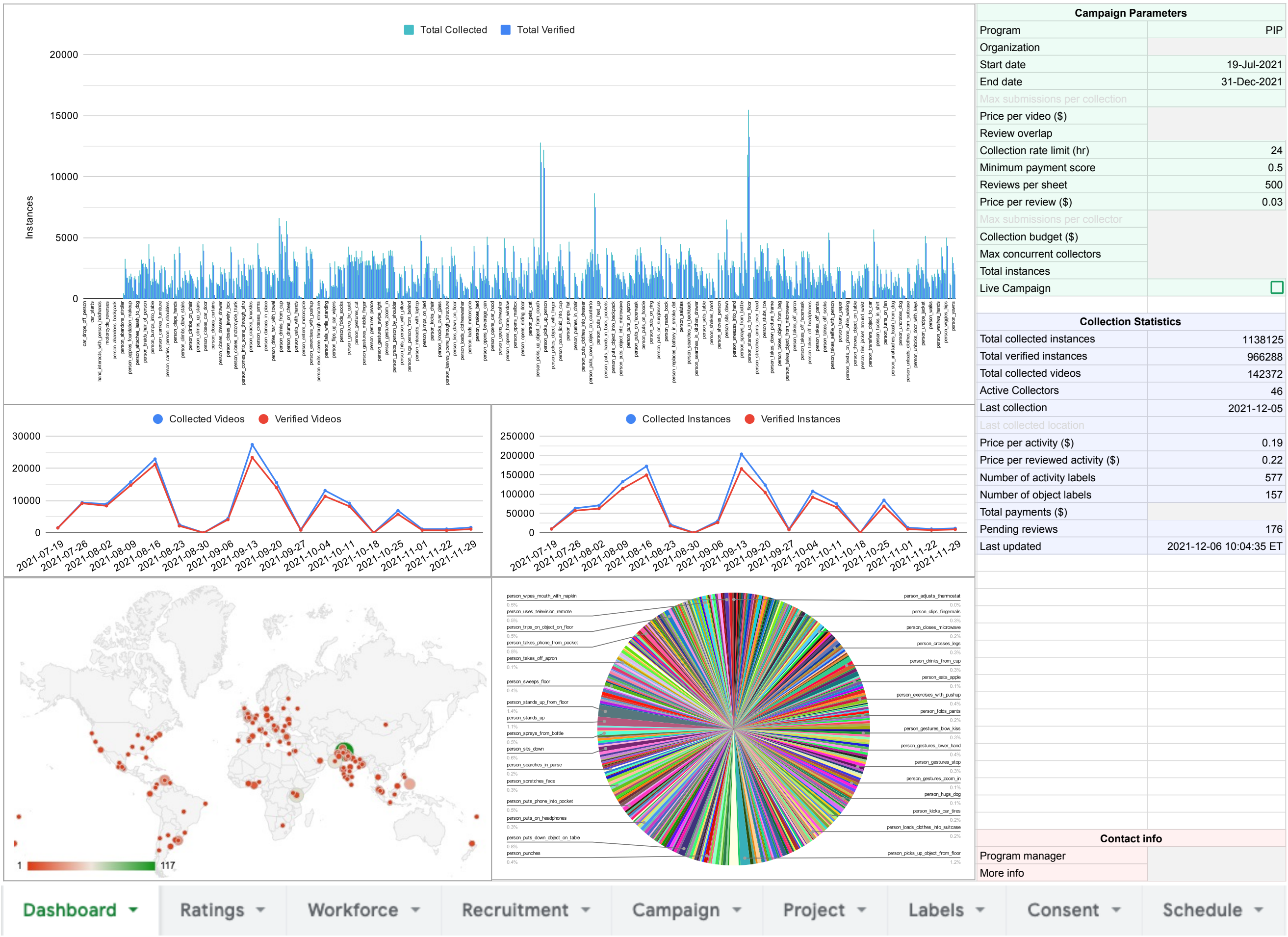
Figure A.11. Editing in the Collector mobile app. The editor is an in-app tool that allows for modifying the bounding box and activity timing for objects in video. This allows for correction of live annotations that were inaccurately collected at recording time. Editing is performed by using multi-touch gestures to define bounding boxes that deform through time to track objects, and press gestures within boxes to define start and end times for activities. A video tutorial for the editor is available at visym.com/editor.

subtle differences in how the lower body is moved in sitting by resting your body weight against a flat surface as compared to squatting down by bending your knees while resting your body weight on your feet. In the second case, there are clear motion differences between drinking with your upper body as compared to sitting with your lower body. This suggests that a fine-grained class requires subtle motion differences with other closely related verbs.

Is a fine-grained activity defined by the noun performing it? For example, compare the activity category of *dog sits* vs. *person sits*. The kinematics of a four legged animal sitting on hind legs exhibits a different motion than a two legged person sitting in a chair. Furthermore, the object appearance of the noun “dog” can provide context to aid in the recognition of *dog sits* as compared to *person sits*. The discrimination of a fine-grained activity class should primarily require representation of the motion being performed and not exploiting object appearance or scene context cues. This suggests that the set of fine grained activities should be performed by the same noun in order to remove the confounding effects of object or scene context. This does introduce a challenge of combinatorial scale when composing noun/verb pairs, however a dataset focused on people only will avoid this combinatorial explosion.

Is a fine-grained activity defined by an object being interacted with? There exist scenarios with an actor interacting with visually distinct objects that exhibit the same or different motion pattern when performing the activity. For example, consider the activities *person throws baseball* vs. *person throws rock*. These activities exhibit largely the same throwing motion of either a rock or a baseball and the use of the object does not change the motion of the activity. These motions are not visually distinct, and are only distinguishable through identification of the object category “rock” or “baseball”. However, consider *person carries bicycle* vs. *person carries groceries*. Carrying a heavy bicycle is awkward and requires a different strategy of carrying over your shoulder or pulling the object towards your chest, as compared to lifting grocery bags with handles in either hand. These are both examples of carrying a heavy object lifting using your upper body then walking, but the motion induced by the object when performing the activity is different. This suggests that a set of fine-grained activities should include object interactions that induce visually distinct motions.

Is a fine-grained activity defined by the style in which it is performed? An activity style can be described in terms of an adverb that modifies the verb being performed, such as “skillful” or “clumsy”. We humans are experts at subtle discrimination between gestures or social interactions and we are highly tuned to picking up on the body language in


Total Collected
Total Verified
Instances
20000
15000
10000
5000
0
Collected Videos
Verified Videos
30000
20000
10000
0
2021-07-19
2021-07-26
2021-08-02
2021-08-09
2021-08-16
2021-08-23
2021-08-30
2021-09-06
2021-09-13
2021-09-20
2021-09-27
2021-10-04
2021-10-11
2021-10-18
2021-10-25
2021-11-01
2021-11-22
2021-11-29
Collected Instances
Verified Instances
250000
200000
150000
100000
50000
0
1
117
person_wipes_mouth_with_napkin 0.5%
person_uses_television_remote 0.5%
person_trips_on_object_on_floor 0.5%
person_takes_phone_from_pocket 0.5%
person_takes_off_apron 0.1%
person_sweeps_floor 0.4%
person_stands_up_from_floor 1.4%
person_stands_up 1.1%
person_sprays_from_bottle 0.5%
person_sits_down 0.6%
person_searches_in_purse 0.2%
person_scratches_face 0.3%
person_puts_phone_into_pocket 0.5%
person_puts_on_headphones 0.3%
person_puts_down_object_on_table 0.8%
person_punches 0.4%
person_adjusts_thermostat 0.0%
person_clips_fingernails 0.3%
person_closes_microwave 0.2%
person_crosses_legs 0.3%
person_drinks_from_cup 0.3%
person_eats_apple 0.1%
person_exercises_with_pushup 0.4%
person_folds_pants 0.2%
person_gestures_blow_kiss 0.3%
person_gestures_lower_hand 0.4%
person_gestures_stop 0.3%
person_gestures_zoom_in 0.1%
person_hugs_dog 0.1%
person_kicks_car_tires 0.2%
person_loads_clothes_into_suitcase 0.2%
person_picks_up_object_from_floor 1.2%
Dashboard
Ratings
Workforce
Recruitment
Campaign
Project
Labels
Consent
Schedule

Campaign Parameters	
Program	PIP
Organization	
Start date	19-Jul-2021
End date	31-Dec-2021
Max submissions per collection	
Price per video ($)	
Review overlap	
Collection rate limit (hr)	24
Minimum payment score	0.5
Reviews per sheet	500
Price per review ($)	0.03
Max submissions per collector	
Collection budget ($)	
Max concurrent collectors	
Total instances	
Live Campaign	☐

Collection Statistics	
Total collected instances	1138125
Total verified instances	966288
Total collected videos	142372
Active Collectors	46
Last collection	2021-12-05
Last collected location	
Price per activity ($)	0.19
Price per reviewed activity ($)	0.22
Number of activity labels	577
Number of object labels	157
Total payments ($)	
Pending reviews	176
Last updated	2021-12-06 10:04:35 ET

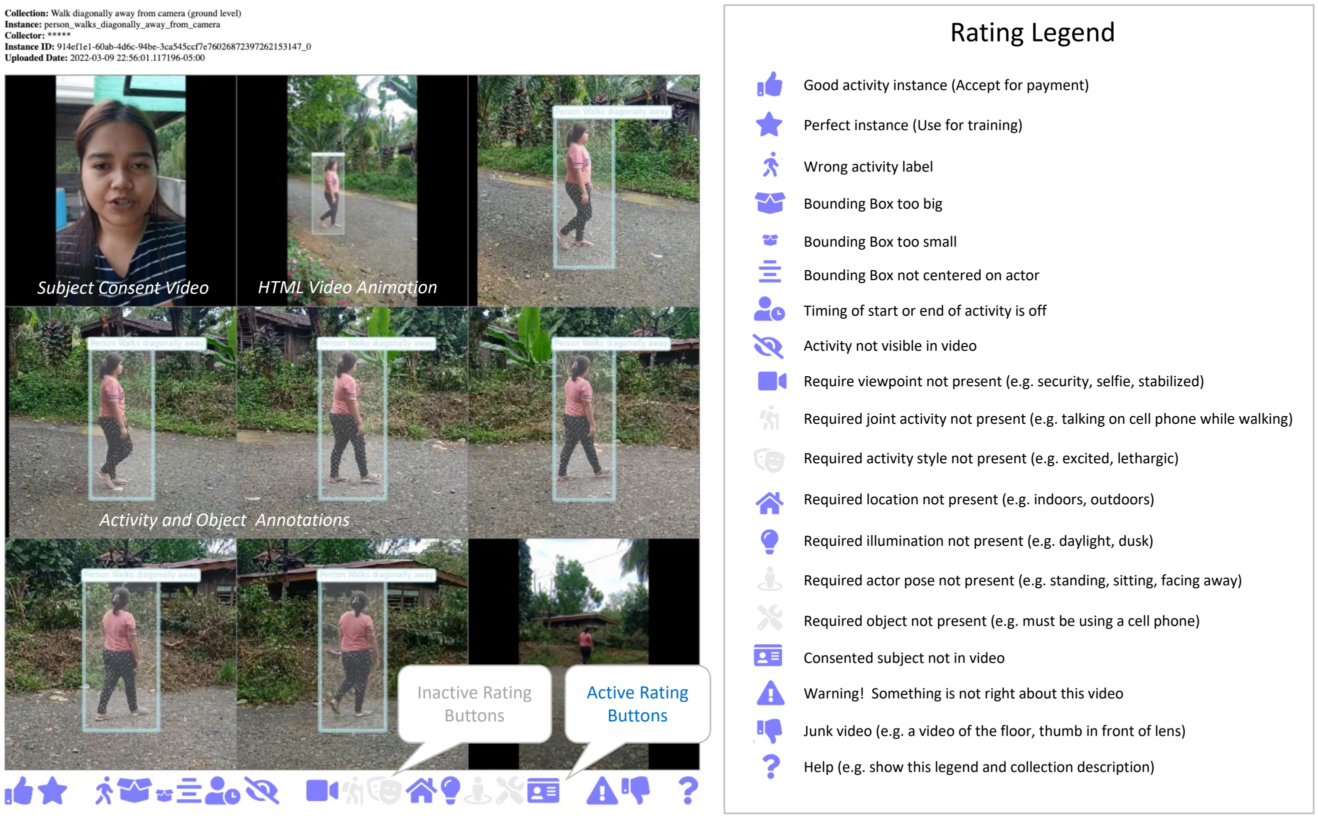
Contact info	
Program manager	
More info	

Figure A.12. Campaign Dashboard on the Collector platform. The campaign dashboard is a live HTML interface used by a campaign administrator to monitor collection status, view live submission stream, list active collectors worldwide, modify collections in-app, define new labels and specify IRB consent language. The bottom tabs expand into specific views of the campaign focused on Ratings, Workforce payments, Recruitment onboarding/offboarding and Campaign specification. The dashboard shows the live histogram of total collected instances, collected and accepted videos by week, submissions by geographic distribution, and pie graph of labels submitted and total collection statistics. Finally, the campaign parameters provide command and control of the collection campaign.

how an activity is performed. However, the same visually grounded style can be performed for more than one activity, such as *person sits skillfully* as a dancer would or *person jumps skillfully* using an efficient economy of motion. This suggests that style is an important attribute for within-class variation, but should not be considered as a separate class.

Finally, why do we need atomic fine-grained activities? Our core hypothesis is that training a visual AI system that represents the diversity of human activities will result in a representation of subtle motions with improved generalization performance for real scenes. Large-scale pretraining requires significant overlap between label set and target domain [55][11]. Our collection goals are to collect the activities of daily life from third person viewpoints to provide a pretraining and fine-tuning dataset for activity detection of daily life. For example, to detect when a person is talking on a cell phone, we can include closely related activities in training such as scratching your head or putting on headphones. This forces a more precise learned representation of talking on a cell phone that is not always predicted simply when you touch your ear. Furthermore, a training dataset that includes simple, short and atomic activities can provide a foundation for study of longer complex or composite activities of daily life that require reasoning about intent or identity. Our goal is to explore this representational ability for atomic activities, and provide a foundational open dataset to explore ethical human activity detection.

Long-tailed classes. Human activities are diverse and long-tailed. There exist activities that each of us perform many times per day, such as standing up or sitting down, opening or closing doors or getting dressed. These common activities are diverse in that there are many ways that one can



Collection: Walk diagonally away from camera (ground level)
Instance: person_walks_diagonally_away_from_camera
Collector: *****
Instance ID: 914ef1e1-60ab-4d6c-94be-3ca545ccf7e76026872397262153147_0
Uploaded Date: 2022-03-09 22:56:01.117196-05:00

Rating Legend

- Good activity instance (Accept for payment)
- Perfect instance (Use for training)
- Wrong activity label
- Bounding Box too big
- Bounding Box too small
- Bounding Box not centered on actor
- Timing of start or end of activity is off
- Activity not visible in video
- Require viewpoint not present (e.g. security, selfie, stabilized)
- Required joint activity not present (e.g. talking on cell phone while walking)
- Required activity style not present (e.g. excited, lethargic)
- Required location not present (e.g. indoors, outdoors)
- Required illumination not present (e.g. daylight, dusk)
- Required actor pose not present (e.g. standing, sitting, facing away)
- Required object not present (e.g. must be using a cell phone)
- Consented subject not in video
- Warning! Something is not right about this video
- Junk video (e.g. a video of the floor, thumb in front of lens)
- Help (e.g. show this legend and collection description)

Figure A.13. Reviewing interface on the Collector platform. Reviewers are provided an HTML interface for each video under review, that provides buttons for feedback. The reviewer can click on the image to show a WEBP animation of annotated video, along with the consent video in the upper left to confirm that the recorded subject is the subject that provided consent. Reviewers are tasked with selecting one of the review buttons at the bottom for this video, which are streamed to the Collector backend for aggregation. The rating legend shows the description for each buttons, such that "grey" is disabled and is not required for this submission.

perform each activity which change the appearance, such as sitting criss-cross on the floor vs. sitting in a chair, opening a sliding glass door vs. opening a facility door or putting on a hat vs. putting on a jacket. These within-class variations of activities are in addition to the more common variations due to camera pose, actor pose or illumination. These within-class variations of activities are *diverse* which capture the variability of naturally occurring human activities.

Human activities are also *long-tailed*. Just as there are activities that each of us perform frequently, there are many more activities that we perform infrequently or possibly never. For example, for some people this may be domestic activities such as cooking, cleaning or folding, for others it may be violent activities, such as fighting, others may be potentially harmful activities such as tripping or falling. These are activities that may occur so rarely that in months of video (or scraping videos from social media) no examples are captured. Furthermore, even if these activities occur, annotating them in long duration videos requires manual search through many hours of video to localize rare activity instances. It becomes increasingly difficult for annotators to remember all activities that have been specified as the number of activity classes increases. This imbalanced frequency distribution of the occurrence of human activities is long tailed in that there are fewer classes that are performed frequently, but many more classes in the tail of this distribution that are infrequent.

Importance sampling. How can we create a balanced dataset that includes both diverse and long-tailed human activities? One strategy is to consider a video as a sample of the visual world, such that a video dataset contains a finite sample for a specific task. Direct sampling (e.g. point a camera out the window) leads to a dataset that may be large scale, but imbalanced, containing frequently occurring labels (e.g. people walking), but will under-represent rare and fine activities that may never occur here.

Consider an alternative strategy of *importance sampling*, which samples videos given pre-selected labels. This strategy generating samples on-demand for nearly any desired label.This introduces a tradeoff between direct sampling which curates videos that are naturally occurring and frequent (but imbalanced) vs. importance sampling that are

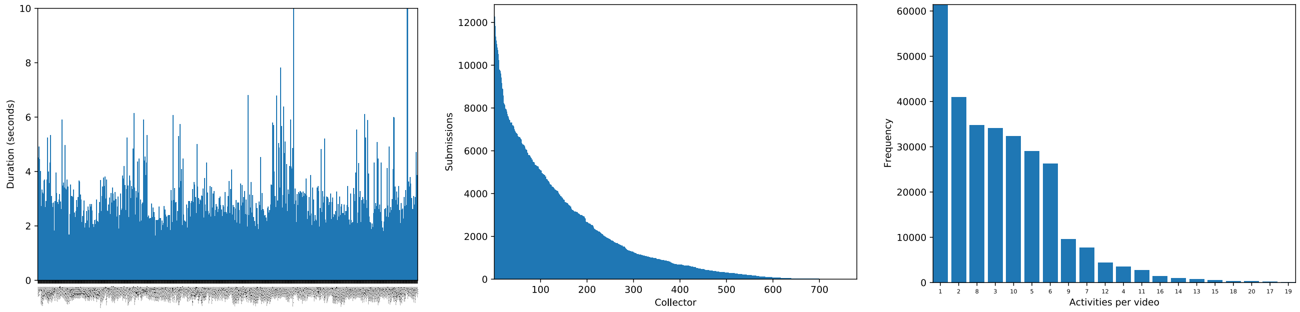


Figure B.14. CAP Dataset statistics. (left) Mean duration in seconds of each activity label, (middle) The number of submissions for each collector sorted by maximum number of submissions which shows that some collectors are enthusiastic about making contributions, (right) The sorted density of the number of activity instances per video, which shows that the most common number of activity instances per video are 1, 2 or 8 instances. We recommend zooming into the PDF to see the activity labels.

engineered and balanced (but potentially biased).

Section 3 introduced the Collector platform to address the issue of diverse and long-tailed classes by enabling *on-demand collection*. The Collector mobile app enables activities to be requested, recorded, annotated and submitted on-demand from a global workforce of collectors. These collection requests are controlled to balance out the frequency of long-tailed classes. Furthermore, the collections break down activities into variations that capture the diversity of the activity class. For example, to address the diversity of "person dressing" we can release a collection request for different variations of dressing: person puts on hat, person puts on shirt, person puts on belt, person puts on pants, person puts on socks, person puts on shoes, etc. To address the diversity of illuminations, we can request a video in indoor illumination conditions or in darkened conditions. This platform operationalizes the dataset collection strategy of importance sampling. Section B.4 discusses this further.

Domain Adjacency. Domain adjacency is the problem of collecting training and testing data in a given source domain, which will be deployed to a closely related (but not identical) target domain. An ideal machine learning system will be evaluated on the same domain data used for training to maximize performance, however there exist domains for which collection of source test data is prohibited by cost, ethical restrictions or policy. This introduces a domain shift between training and testing that will affect performance, and must be mitigated by fine-tuning on source domain data or domain adaptation.

Visual AI has the potential to give us helpful and personalized insights into the rich patterns of our daily life. However, we humans spend the majority of our time in what may be called the *dark domains* of visual AI. These are locations or data sources that are not broadly exposed to visual AI today due to privacy concerns, robustness issues or a lack of datasets. For example, third person viewpoints of people in private or shared spaces could enable helpful applications of ambient healthcare [31][56], wellness monitoring [4][30] or ethical security [14][65][34][7]. However, these spaces contain our private data, and privacy regulations (e.g. GDPR, BIPA, CPRA) require that visual AI address collection and protection of private data by design.

Consider the collection of long duration videos from a third-person viewpoint in private spaces. These videos may require hundreds or thousands of hours per camera to capture rarely occurring activities, all while watching and recording the intimate details of the private lives of subjects. This introduces (i) an engineering challenge of sharing this huge volume of data, (ii) a sparsity challenge of efficiently sharing data that is mostly empty video between interesting activities, (iii) a long-tailed challenge of collecting data of rare activities that may never occur organically, (iv) a bias challenge where people who have consented to recording and know they are being recorded change their behavior [4] and (v) an ethical challenge of whether we should share data from private spaces in the first place.

This suggests that direct collection and distribution of in-domain data for these dark domains should be avoided. There exists a tradeoff between the volume of data curated vs. the quality of the data collected for target domain deployment. Our fundamental hypothesis is that collecting a large amount of annotated data in a closely related source domain (e.g. short duration, on-demand, third-person videos), then deploying the trained system to an adjacent target domain with privacy restrictions (e.g. long duration third-person videos in private spaces) will enable ethical deployment of a trained system for dark domains. However, a key challenge is collecting domain adjacent data for these dark domains without compromising visual AI system performance. This issue will be explored in section B.7.

B.4. Dataset Collection

Dataset collection is the process of defining, recording, annotating, verifying and distributing a dataset that achieves

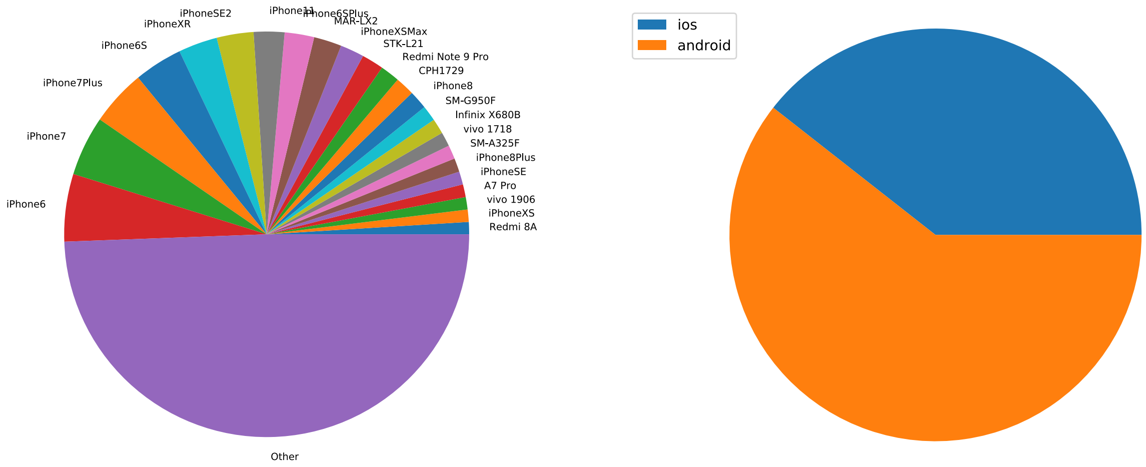


Figure B.15. CAP mobile devices. Distribution of mobile devices and mobile OS from collector submissions. This shows that there is a wide variety of mobile devices used to collect the CAP dataset, with a slight preference for Android devices.

the design goals in section B.2 while addressing the key challenges in section B.3. The dataset collection leverages the Collector platform introduced in section 3 to collect the data. In this section, we address design issues related to setting up the Collector campaign.

Collection Campaign. The collection campaign was specified on the Collector platform as follows:

- **Collections**. The campaign specification includes 842 unique collection types, each specifies one of 512 activity labels with person interaction or interactions with 157 object types. 288 of 842 collection types were specified to be collected so that the subject is "facing away" from the camera, so that the back of the subject is more visible than the front, such that the activity may be partially occluded to increase diversity.
- **Repetitions**. Collectors were tasked with repeating activities between 5-10 times in each submission. This provides in-video data augmentation, but rather than synthetic augmentations (e.g. mirror, crop, scale). We ask collectors to perform the activity slightly differently each time, enabling natural data augmentation.
- **Third-person viewpoint**. Collectors were tasked with collecting all videos from a third-person viewpoint, looking slightly down on the scene from above, to model a security camera on the wall or ceiling.
- **Physically stabilized**. 38 of 842 collection types were specified to be physically stabilized. We tasked collectors to rigidly mount their device to a simple triangular cardboard stand and put it on a flat surface up high looking down (e.g. a shelf, a stack of boxes, a ladder). Collectors record their subject (or themselves) performing the scenario, then they use the in-app editor to annotate when and where the activities were performed.
- **Temporal activity detection**. 87 of 842 collections were specified to be collected to support temporal activity detection. We instruct the collectors to choose from a list of 11 activities to perform in a natural sequence called a "scenario". For example, one scenario is "Come inside from the cold" which includes the activities of entering a room and taking off winter outerwear. The subject performs at most 11 activities in any order they choose, then the collector records the scenario, and edits in-app to annotate the performed activities.

The overall collection statistics are shown in figure 2 which shows the submission locations of collectors worldwide, along with the label histogram in figure 4. Figure B.14 describes additional supporting details about the dataset including collector submission frequency, bounding box size distribution (Figure B.16), mean duration per collection type and activity density per video, and mobile OS and device type distribution (Figure B.15).

Finally, a visualization of the scale of this dataset is shown in figure B.19, which shows a montage of less than 1% of the videos, along with a visualization tool to interactively explore the dataset. Figure B.21 shows a montage of ground truth examples of the activity detection task. This shows samples of frames from the activity detection task that shows the sequences of activities that a collector is tasked to perform in eight scenarios.

B.5. Dataset Format

The CAP dataset is publicly available for download. In this section, we provide details on the dataset format, including the selection of the label naming format, post processing for bounding box improvement and background stabilization and additional video metadata.

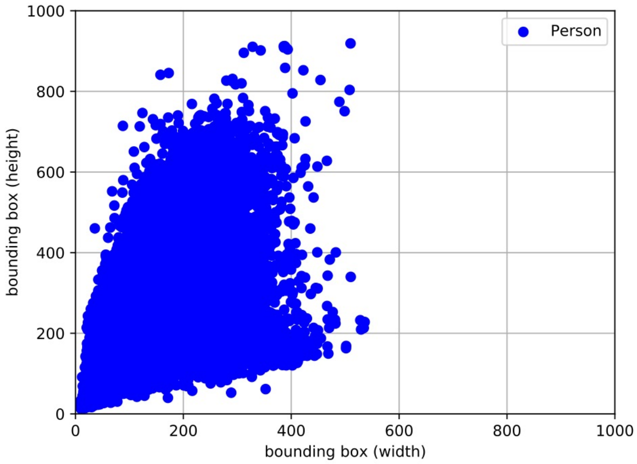


Figure B.16. CAP 2D Bounding Box Distribution showing the size variation of labeled people boxes.

Activity Label Nomenclature. The label format for atomic activity labels follows the following compositional structure: "noun verb adjective noun". For example, *person opens car door* or *person opens refrigerator door*. This structure provides a consistent naming structure for atomic activities, that allows unambiguous description of an atomic activity with optional adjective extensions to provide specificity for person-object or person-person interactions. Furthermore, we specify the hierarchical label structure using only the "noun verb" components of this label, with the remaining label components specifying the within-class or fine-grained variation.

The label format for this dataset is in contrast with other large scale activity datasets. For example, the caption label style of Charades [72] or Something-Something v2 [27] describes a label as a phrase or short natural language narrative of the contents of the video (e.g. Putting a book somewhere, Approaching something with your camera). The verb only label style of Moments in Time [59], AVA [28], HMDB [46], Kinetics [41] and ActivityNet [8] describes an activity in terms of the verb being performed, largely independent of the object being interacted with or actor performing the verb. Finally, our nomenclature goal is to provide a more intuitive label description than alternatives previously deployed, such as wordnet synsets [18]. Our goal is to provide more specific representation of the within class variation of activity classes, by exploring the actor, person-object and person-person interactions as represented in the label name. This label nomenclature is closely related to the Multiview Extended Video with Activities (MEVA) class naming [14].

The label format for this dataset is also in contrast with *open vocabulary datasets* [20][29]. In this style of dataset collection, raw data is recorded in a target domain explicitly without a target task in mind for this data. The raw data is collected, then it is post processed by an annotation team to provide labels or natural language captions for the data that was collected for a task defined after collection was performed. For example, the Ego4D dataset [20] recorded egocentric video from first person perspective of wearers going about their daily lives, which is captioned after collection. This data provides a sample of the common activities that were performed during the recording period, but this does not provide the training data needed for supervised learning. Similarly, the LVIS dataset [29] for large vocabulary long tailed object recognition collects cluttered images in natural settings and asks annotators to achieve consensus for labeling all of the objects that are present in these images. This can achieve dense labels in naturally occurring images, however it cannot achieve balanced datasets.

Weak Annotation. Videos are post-processed after collection to include an optional step of weakly annotated bounding box refinement. Weak annotation refinement is the process improving the bounding box provided in-app by the collector. The annotation by the collector provides a weak label that is collected quickly and easily, which coarsely overlaps the true object, with minor misalignment errors. Then, a pretrained object detector selects an optimal low confidence proposal that maximizes overlap and confidence with the weak annotation. This strategy significantly reduces the cost of large scale annotation by performing the annotation while recording videos, with error correction in post-processing.

The weak annotation refinement is performed as follows. Videos are processed with a low confidence object detector for the target actor in the video (e.g. a person detector) forming low confidence object proposals. Note that this strategy requires a pretrained object detector for the target class. Next, proposals per frame are grouped using maximum intersection over union (IoU) assignment forming object tracks for each object instance in the video. Next, given a sequence of object bounding boxes from a human annotator, object tracks are rescored to compute the framewise product of proposal confidence and IoU with the annotated box, followed by a mean score over all frames in the track. Finally, the object tracks are sorted by this rescored confidence, and the track with the highest score selected as the weak annotation refinement. This selects the object track that maximally overlaps the weak annotation with highest confidence, forming a weak annotation refinement. This procedure assumes that the annotated box from the collector overlaps the primary actor by at least 50%, which is enforced during the human review process. Both the refined box and the collected box are exported in the dataset release. The weak annotation code is open source.

Figure B.17 shows an example of the weak annotation refinement. In this montage, each montage element is a frame from a collected video. We show the bounding box

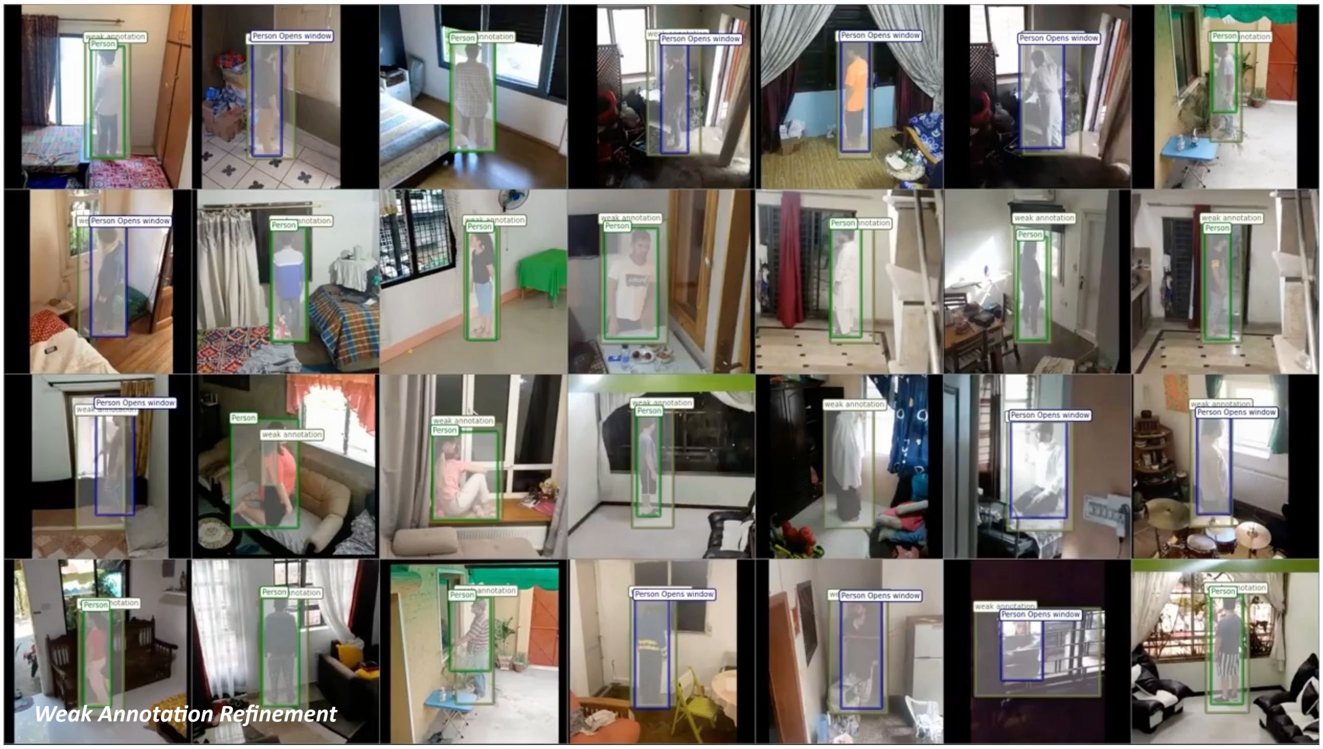


Figure B.17. Dataset post-processing by weak annotation refinement. The human annotation (captioned "weak annotation") shows the box defined by the collector in-app while recording, which is generally centered on the subject, but may not be tight around the limbs. The refinement (captioned "Person") selects an person track using object detection proposals that optimally overlaps with the weak annotation. Both annotations are available in the public dataset release.

annotated by the collector live during collection with the caption "weak annotation". This provides the weak label which is improved into the refined bounding box with the caption "Person" (or "Person Opens Window" if the subject was in the process of performing this activity in this video). This shows that the refinement procedure creates boxes that are tight around the torso and limbs, even when the subject is in a atypical pose or partially occluded in the scene. However, this derived method can still introduce rare assignment errors (e.g. row 3 column 2), or missing annotations if there is no overlap between tracks and annotations, so the end-user should be aware of possible corner cases and fall back on the weak annotations as needed. The weak annotated refinement is best shown in a video, showing weakly annotated bounding boxes sampled from the CAP dataset.

Background Stabilization. Background stabilization is the process of stabilizing the camera to the first frame so that the background is unmoving. This strategy reduces the cost of large scale dataset collection by not requiring that the cameras be rigidly mounted on a tripod, since few freelancers have tripods with mobile device mounts available for use. The collector can record videos handheld, which are post-processed to stabilize the videos as if they were rigidly mounted and unmoving.

The approach for background stabilization is affine stabilization to frame zero using multi-scale optical flow correspondence with foreground object keepouts. This pipeline supports optical flow based stabilization of video which reduces the artifacts due to hand-held cameras to stabilize the background. Remaining artifacts are due to non-planar scenes, rolling shutter distortion and subpixel optical flow correspondence errors. The stabilization is only valid within the tracked actor bounding box for small camera motions. Large motions will introduce stabilization artifacts due to non-planar scene effects and should be filtered prior to usage. The stabilization artifacts will manifest as a slightly shifting background relative to the actor which may affect flow based methods. Finally, the approach transforms all bounding boxes to be aligned with the stabilized video, and includes the stabilization residual in video metadata to enable filtering stabilization with poor alignment. The stabilization code is open source.

Figure B.18 shows an example of the background stabilization. The background stabilization is best shown in video, comparing unstabilized handheld video collected from mobile devices to 5Hz background stabilized video. Observe that the background stabilized video in these YouTube links has the background unmoving as if the video

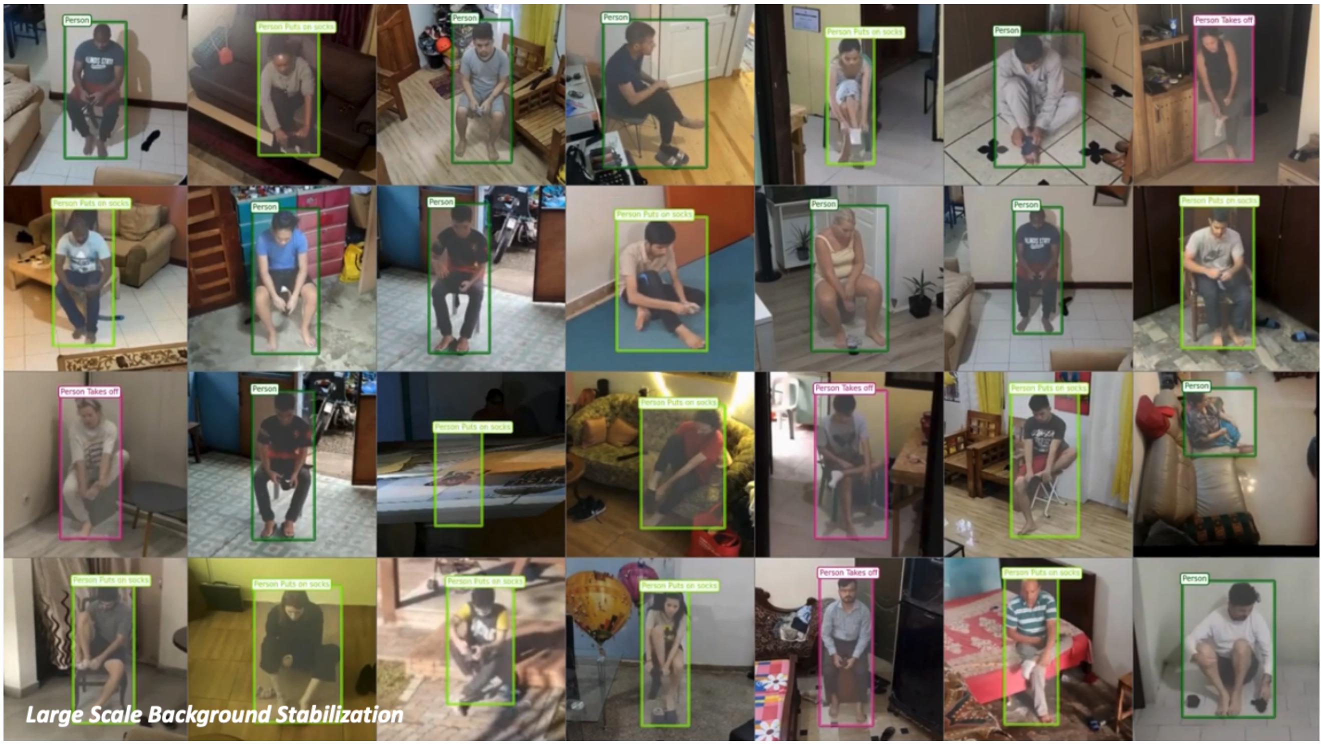


Figure B.18. Dataset post-processing by background stabilization. We use a flow based affine stabilization method to align each frame to the first frame of the video to enforce that that the background is unmoving, as if the video was collected from a rigidly mounted camera. The background stabilization is best shown in video, comparing unstabilized handheld video collected from mobile devices to background stabilized video. Bounding boxes are affine transformed using the affine stabilization transformation to align boxes to the stabilized video.

was collected from a rigidly mounted static camera.

Privacy and Consent. We require that all subjects review a consent form, and provide their informed consent for the dataset collection. This consent form is Institutional Review Board (IRB) approved and describes how the data will be collected, what data will be collected, how it will be shared and who it will be shared with. The all subjects consent to their likeness to be shared in publication material and a publicly accessible dataset release for the purposes of visual AI research. Each subject is required to provide a video consent, which is a selfie video of this subject stating that they consent to the video collection. This selfie video is used by the review team to compare that the subject recorded in the video is the consented subject. This enables rejection of fraudulent video submissions of non-consented subjects. An example of this selfie consent video is shown in the upper left of the review interface in figure A.13. IRB consent forms shown in-app are customizable for local IRBs.

Non-consented subjects in the video field of view have their faces blurred out in-app prior to submission. All subjects with visible personally identifiable information (PII) in the videos have consented to having their PII shared for the purposes of visual AI research. Non-consented subjects are those subjects that are not within the foreground box of the primary actor. All non-consented subjects have their faces blurred out on-device in the mobile app prior to submission. This feature is currently enabled only in the iOS release.

Collection Metadata. The collection platform includes the following additional metadata released for each video:

- **Program name/ID.** This is the name of the campaign under collection. For the CAP dataset, the program name was either PIP or MEVA. Program names (and associated program ID) are globally unique on the collector platform.
- **Collection name/ID.** This is the description of the collection that is given to the collectors. Collection names are globally unique to a program. The ID is a globally unique string that uniquely identifies this collection.
- **Collection date.** This is the timestamp when the video was collected in the local timezone of the device.
- **Geolocation.** This is the region of the world that the collection was recorded (with consent). The geolocation is captured with the field "ipAddress" which is the public IP address of the internet service provider of the device. IP geolocation services can be used to convert this IP address to a geographic region.
- **Collector ID.** A globally unique identifier for the reg-

istered collector who was logged into the mobile app and is recording the video. Collector IDs have been anonymized to avoid association with email addresses.

- **Subject ID.** A globally unique identifier for the consented subject who is in the video. This is the ID of the subject who was identified at consent time.
- **Mobile device.** The mobile OS (Android, iOS) and device hardware that was used to collect the source video.
- **Frame rate.** The frame rate in fractional frames per second at which the original video was collected.
- **Video dimensions.** The video (height, width) in pixels.
- **Rotation.** The rotation state (e.g. landscape, portrait, left, right) of the device when recording.
- **Blurred faces.** The number of non-consented faces blurred on-device prior to video submission. This feature is enabled for iOS devices only. Non-consented faces are those not within the collector annotation box.
- **App version.** A version string that identifies the release version of the mobile app used to collect this video.
- **Video ID.** A globally unique ID that uniquely identifies this video on the collector platform.

Bias Engineering. Our collections organize activities into groups to introduce diversity in the scene. For example, we specify to the collectors to load and unload both from a trunk and from a rear door of a vehicle to help introduce within-class diversity. Also, we introduce joint activities such as "Leave this scene while talking on a phone". Finally, we specify that collectors should be facing or facing away from the camera to introduce more pose diversity. The full list of collection names are self explanatory and may be filtered to remove variants that may not reflect the target domain bias, or which do not satisfy the assumptions of the loss function of the target system.

Temporal Padding. All data is distributed in a clipped or padded form. The clipped dataset includes activities that are tightly temporally cropped around each activity, such that the duration of the clip is the duration of the video. The padded dataset temporally pads each video to $\geq$ 3s, along with the metadata to recover the tight clip. This provides additional video context for each training video to support the temporal assumptions of a target system or additional data augmentation.

Recommended Splits. The training, validation and test set splits are generated using an 80/10/10 split strategy controlling for collector ID. The dominant source of bias in the dataset is the effect of collectors submitting different activities in closely related locations, such as the same house or yard. We seek to avoid the same collector providing videos for both training and testing, in order to create a more realistic testing scenario. To achieve this, we randomly split the collector IDs into 80/10/10 splits, then assign all videos submitted by this collector into the corresponding training/validation/test sets. The test set is sequestered for leaderboard evaluation purposes, and videos are available for download behind a license agreement restricting redistribution.

Figure B.14 shows the distribution of videos per collector which shows a falloff in submission frequency and allows a random sample of collectors to avoid unbalanced video distribution in the dataset. The final split enforces that collector IDs are disjoint between training, validation and test. The splits are included in the release metadata.

Annotation format. Dataset management includes processing annotations and pixels for a large number of videos. In order to streamline this data processing pipeline, we have developed the open source vipy package. Vipy is a Python package for representation, transformation and visualization of annotated videos and images. Vipy provides tools to apply transformations such as downsampling, padding, scaling, cropping and rotating so that the annotations are transformed along with the pixels. The vipy annotation format is open JSON designed for representation of activity and object annotations in video.

B.6. Benchmark Research Questions

This dataset provides the data to answer the following research questions. The answers to these research questions are provided in section 6.1.

Fine grained categories. Is there an improvement when training using fine grained categories (e.g. *person picks up object from floor* vs. *person picks up object from table*) vs. coarse grained categories (e.g. *person picks up object*) when testing coarse grained activity classification and temporal activity detection?

Collection diversity. Is there an improvement when training with explicitly engineered within-class diversity (e.g. two biases are are controlling for are activity instances collected with actor pose explicitly "facing away" from the camera, and instances collected in a crowded scene with nearby occluding people).

Collector diversity. What is the relationship between the number of unique collectors in the training set vs. test set performance? How many unique collectors do we need? Does it help to have one collector doing many collections in the same location and clothes?

Stabilization. What is the effect when training software background stabilization from handheld collection on rigidly mounted videos? Can we correct for this domain shift through software stabilization?

Video data augmentation. Is there a benefit using actor data augmentation (e.g. collectors repeating activities slightly differently each time) vs. synthetic data augmentation (e.g. crops, scales, rotations)?

B.7. Benchmark Evaluation

Performance benchmarking is the specification of an evaluation methodology, task and dataset along with a baseline system design to evaluate system performance. However, section B.3 discussed that this is impractical for long duration third person videos. For example, we may collect thousands of hours of video from a security camera without ever collecting an organic instance of *person puts on shoes*. How do we realistically benchmark a task where the labels to evaluate may never occur? Furthermore, even if we did have enough instances of all the target labels, how do we address the ethical concerns of publicly sharing long duration video of the private daily lives of humans without this being interpreted as a fishbowl or worse, exploiting people to create a human zoo?

Section B.3 addressed this key challenge by introducing *domain adjacent benchmarking*. In this strategy, we collect test sets that are from the required viewpoint, but with actors performing the test activities in short bursts, rather than real subjects going about their daily lives. This provides performance evaluation of the dark domain (e.g. third person, long duration videos collected in private spaces) in a closely related adjacent domain (e.g. third person, short duration videos acted in shared spaces). The test data in the adjacent domain can be collected and distributed ethically, and performance evaluation on the domain adjacent data is used as a surrogate for the dark domain.

However, this strategy exposes a fundamental tradeoff between bias and diversity. As discussed in section B.5, the Collector platform controls diversity through *bias engineering*, where we explicitly request collection parameters to encourage diversity of videos submitted. This allows collection of rare activities that may not occur organically, however these subjects know they are being recorded. Their payment depends on their submission passing verification, which can encourage movements in an exaggerated or unnatural manner. This introduces an “acting bias” into the dataset, that would not be present in the target dark domain. More generally, we have observed the following biases when collecting domain adjacent test data:

- **Duration bias.** All videos are limited to a maximum duration of 45 seconds. This duration bound is due to the practical limitations of uploading large videos from cellular data connections in less developed countries, and the design goal of avoid long duration videos where nothing interesting occurs.
- **Actor bias.** People sometimes perform in an exaggerated manner when they know they are being filmed, resulting in awkward or too well-framed scenes.
- **Handheld bias**. 95% of the dataset is collected handheld which leads to moving camera artifacts. Approximately 5% is collected physically stabilized, and all videos are background stabilized in post-processing.
- **Sequence bias**. Activities are often performed in a proscribed sequence. It is difficult for a subject to remember all the things they are supposed to do, so they often perform the same activities in the same order.
- **Center bias**. Actors are always in the video center and never occluded by image boundary. This is due to the in-app annotation methodology where the bounding box is specified by keeping the subject in the box in the center of the camera while recording.
- **Consent bias**. All our subjects are required to consent to using their personally identifiable information for improving visual AI research. As such, we do not have videos of people in large crowds due to the need to get consent from every subject.

Finally, validation of the domain adjacent benchmarking strategy requires an evaluation of the same system on a private (and unreleasable) test set from the dark domain as compared to the public test set in the adjacent domain. This validation is important to establish trust that performance on the domain adjacent test set is predictive of performance in the dark domain, and an initial study was performed in section 6.3 on a subset of the CAP labels. However, additional work is needed to characterize the performance of the full fine-grained label set on security video.

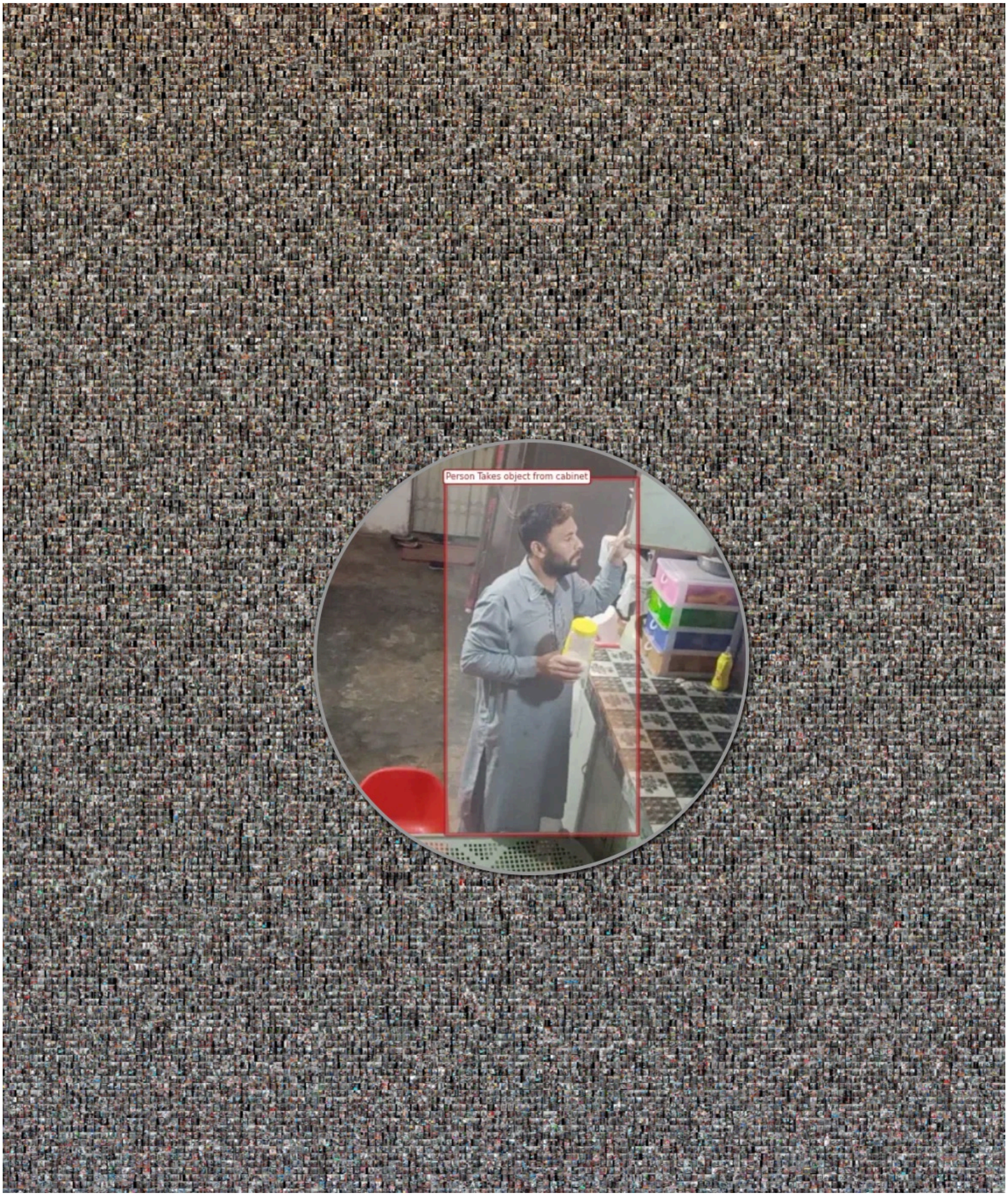


Figure B.19. CAP Dataset Explorer. This visualization shows a 1% sample of the CAP dataset, tightly cropped spatially around the actor and cropped temporally around the fine-grained activity being performed. The full dataset includes the larger spatiotemporal context in each video around the activity, and the complete set of activity labels. This open source visualization tool can hover over a specific video in the montage to show a high resolution animation. The explorer can be sorted by category or color, and shown in full screen.

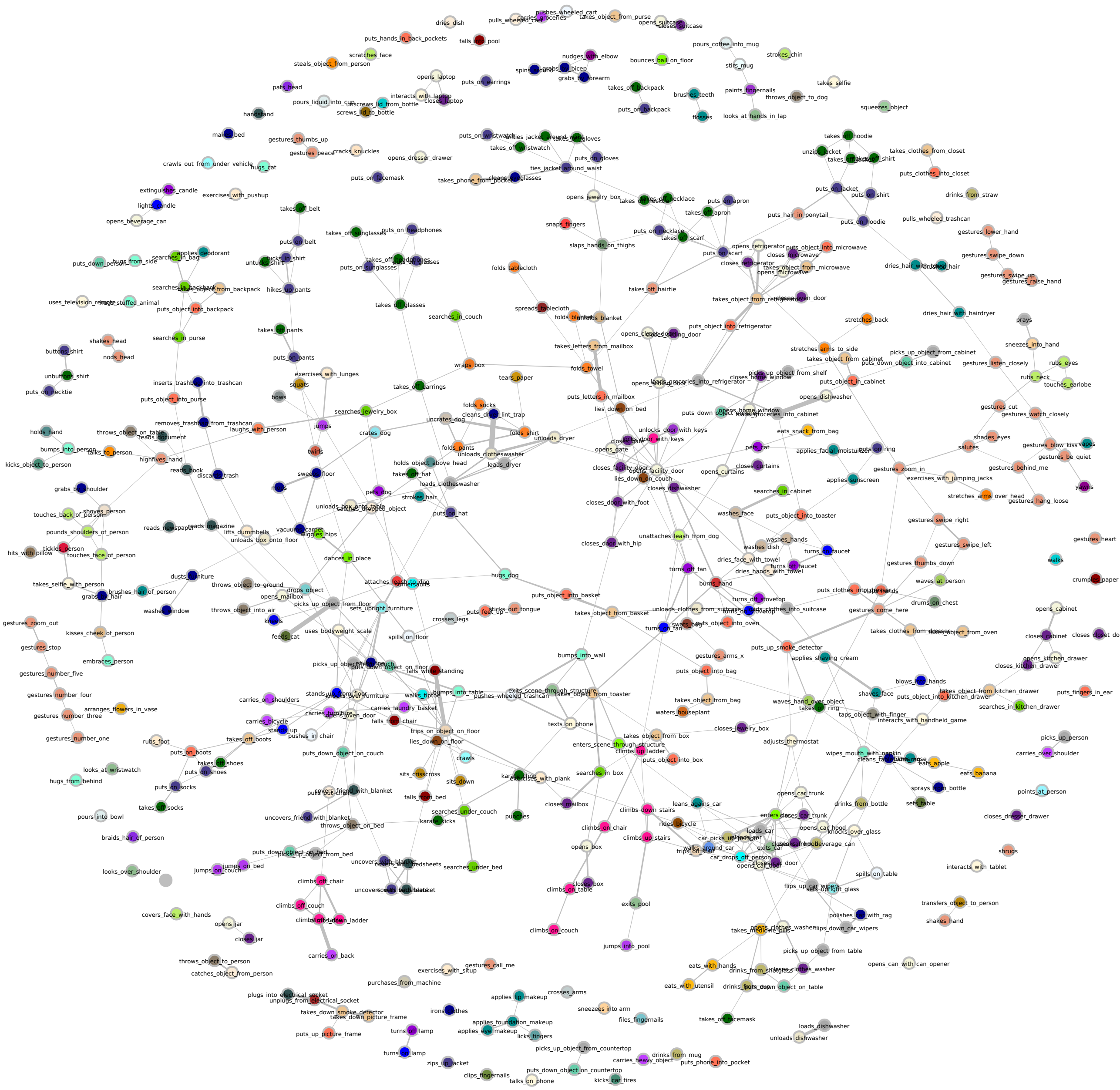


Figure B.20. A confusion graph embedding for the activity classification task in the CAP Benchmark showing edges connecting commonly confused fine-grained activity labels. Node colors correspond to coarse-grained activity labels, and edge thickness corresponds to the confusion weight. The graph embedding was constructed by computing a confusion matrix on the activity classification test set, such that each row of the confusion matrix was normalized to sum to one, then thresholded at 0.07 removing self edges. This forms a sparse adjacency matrix of pairs of labels that are commonly confused. This *confusion graph* visualization provides a 2-d graph embedding of neighborhood structures for commonly confused labels, such that the 2-d graph layout was constructed using a force-directed graph embedding to maintain constant edge length and minimize edge crossings. This provides a visualization of the visual similarity of fine-grained activity classes as compared to the semantic similarity as specified in the coarse-grained label space. We recommend zooming into the PDF to see the node labels.

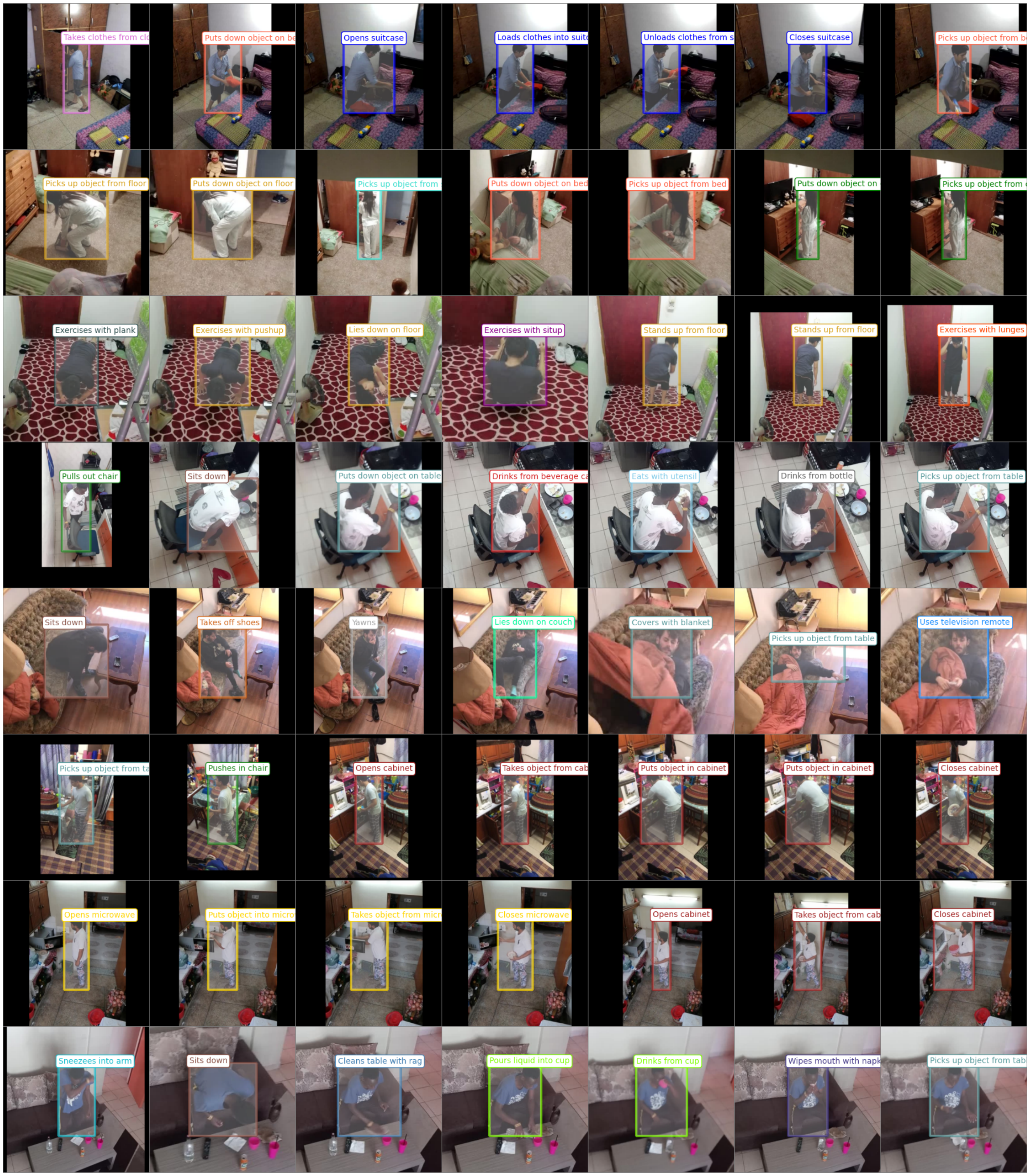


Figure B.21. Activity detection benchmark examples. This visualization shows eight videos sampled from the activity detection task. Each video is collected as a "scenario", which is a sequence of seven to eleven activities performed in any order that the subject in the video chooses. Each row shows the middle frame of the temporal activity annotation, spatially cropped around the primary actor and annotated with an activity caption. The scenarios by row are: "Get ready to go on a trip", "Organize a cluttered room by putting things away where they belong", "Exercise", "Eat a snack", "Sit and watch TV", "Set the dinner table", "Cook a meal in the microwave" and "Drink a glass of water". For example, the scenario "Get ready to go on a trip" includes the activities: take clothes from a closet, put objects onto the bed, open suitcase, load and unload clothes into a suitcase, close a suitcase. A system evaluated on the activity detection task is required to temporally localize these activities in untrimmed clips.

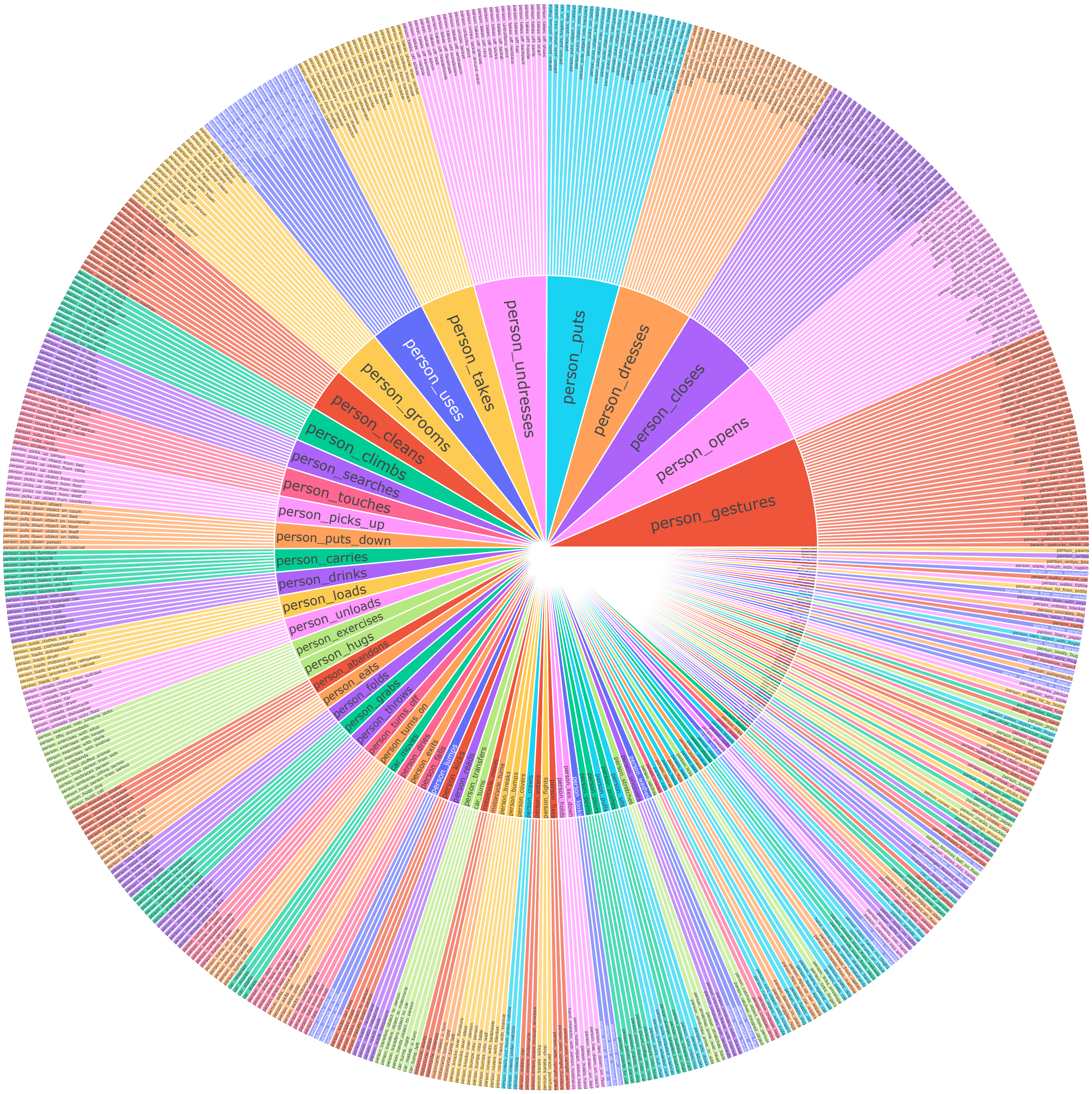


Figure B.22. CAP hierarchical label structure, visualized as a circular tree with outer fine labels grouped by inner coarse labels. We recommend zooming into the PDF to view the labels.