Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives

Kristen Grauman1,2, Andrew Westbury1, Lorenzo Torresani1, Kris Kitani1,3, Jitendra Malik1,4, Triantafyllos Afouras*1, Kumar Ashutosh*1,2, Vijay Baiya*5, Siddhant Bansal*6,7, Bikram Boote*8, Eugene Byrne*1,9, Zach Chavis*10, Joya Chen*11, Feng Cheng*1, Fu-Jen Chu*1, Sean Crane*9, Avijit Dasgupta*7, Jing Dong*5, Maria Escobar*12, Cristhian Furgiua*12, Abrahm Gebreselasie*9, Sanjay Haresh*13, Jing Huang*1, Md Mohaiminul Islam*14, Suyog Jain*1, Rawal Khirodkar*9, Devansh Kukreja*1, Kevin J Liang*1, Jia-Wei Liu*11, Sagnik Majumder*1,2, Yongsen Mao*13, Miguel Martin*1, Efrosyni Mavroudi*7,1, Tushar Nagarajan*1, Francesco Ragusa*15, Santhosh Kumar Ramakrishnan*2, Luigi Seminara*15, Arjun Somayazulu*2, Yale Song*1, Shan Su*16, Zihui Xue*1,2, Edward Zhang*16, Jinxu Zhang*16, Angela Castillo12, Changan Chen2, Xinzhu Fu11, Ryosuke Furuta17, Cristina Gonzalez12, Prince Gupta5, Jiabo Hu18, Yifei Huang17, Yiming Huang16, Wesley Khoo19, Anush Kumar10, Robert Kuo18, Sach Lakshavani5, Mia Liu18, Mi Luo2, Zhengyi Luo3, Brighid Meredith18, Austin Miller18, Oluwatumininu Oguntola14, Xiaqing Pan5, Penny Peng18, Shraman Pramanick20, Meray Ramazanova21, Fiona Ryan22, Wei Shan14, Kiran Somasundaram5, Chenan Song11, Audrey Southerland22, Masatoshi Tatenoko17, Huiyu Wang1, Yuchen Wang19, Takuma Yagi17, Mingfei Yan5, Xitong Yang1, Zecheng Yu17, Shengxin Cindy Zha18, Chen Zhao21, Ziwei Zhao19, Zhifan Zhu6, Jeff Zhuo14, Pablo Arbeláez12, Gedas Bertasius14, Dima Damen16, Jakob Engel15, Giovanni Maria Farinella15, Antonino Furnari15, Bernard Ghanem121, Judy Hoffman122, C. V. Jawahar17, Richard Newcombe15, Hyun Soo Park110, James M. Rehg18, Yoichi Sato117, Manolis Savva113, Jianbo Shi16, Mike Zheng Shou111, and Michael Wray116

1FAIR, Meta, 2University of Texas at Austin, 3Carnegie Mellon University, 4University of California, Berkeley, 5Project Aria, Meta, 6University of Bristol, 7International Institute of Information Technology, Hyderabad, 8University of Illinois, Urbana Champaign, 9Carnegie Mellon University, 10University of Minnesota, 11National University of Singapore, 12University of the Andes, 13Simon Fraser University, 14University of North Carolina, Chapel Hill, 15University of Catania, 16University of Pennsylvania, 17University of Tokyo, 18Meta, 19Indiana University, 20Georgia Tech, 21Project Aria, Meta, 22Georgia Tech

Abstract

We present Ego-Exo4D, a diverse, large-scale multimodal multiview video dataset and benchmark challenge. Ego-Exo4D centers around simultaneously-captured egocentric and exocentric video of skilled human activities (e.g., sports, music, dance, bike repair). 740 participants from 13 cities worldwide performed these activities in 123 different natural scene contexts, yielding long-form captures from 1 to 42 minutes each and 1,286 hours of video combined. The multimodal nature of the dataset is unprecedented: the video is accompanied by multichannel audio, eye gaze, 3D point clouds, camera poses, IMU, and multiple paired language descriptions—including a novel “expert commentary” done by coaches and teachers and tailored to the skilled-activity domain. To push the frontier of first-person video understanding of skilled human activity, we also present a suite of benchmark tasks and their annotations, including fine-grained activity understanding, proficiency estimation, cross-view translation, and 3D hand/body pose. All resources are open sourced to fuel new research in the community.

1. Introduction

A dancer leaps across a stage; Lionel Messi delivers a precise pass; your grandmother prepares her famous dumplings. We observe and seek human skills in a myriad of settings, from the practical (fixing a bike) to the aspirational (dancing beautifully). What would it mean for AI to understand human skills? And what would it take to get there?
Advances in AI understanding of human skill could facilitate many applications. In augmented reality (AR), a person wearing smart glasses could quickly pick up new skills with a virtual AI coach that provides real-time guidance. In robot learning, a robot watching people in its environment could acquire new dexterous manipulation skills with less physical experience. In social networks, new communities could form based on how people share their expertise and complementary skills in video.

We contend that both the egocentric and exocentric viewpoints are critical for capturing human skill. Firstly, the two viewpoints are synergistic. The first-person (ego) perspective captures the details of close-by hand-object interactions and the camera wearer’s attention, whereas the third-person (exo) perspective captures the full body pose and surrounding environment context. See Figure 1. Not coincidentally, instructional or “how-to” videos often alternate between a third-person view of the demonstrator and a close-up view of their near-field demonstration. For example, a chef may describe their approach and the equipment from an exo view, then cut to clips showing their hands manipulating the ingredients and tools from an ego-like view.

Secondly, not only are the ego and exo viewpoints synergistic, but there is a need to translate fluently from one to the other when acquiring skill. For example, imagine watching an expert repair a bike tire, juggle a soccer ball, or fold an origami swan—then mapping their steps to your own body. Cognitive science tells us that even from a very young age we can observe others’ behavior (exo) and map it onto our own (ego) [42, 108], and this actor-observer translation remains the foundation of visual learning.

Realizing this potential, however, is not possible using today’s datasets and learning paradigms. Existing datasets comprised of both ego and exo views (i.e., ego-exo) are few [76, 77, 127, 139, 145], small in scale, lack synchronization across cameras, and/or are too staged or curated to be resilient to the diversity of the real world. Thus the current literature for activity understanding primarily attends to either the ego [28, 47] or exo [48, 67, 105, 149] view, leaving the ability to move fluidly between the first- and third-person perspectives out of reach. Instructional video datasets [103, 159, 204, 207] offer a compelling window into skilled human activity, but (like the above) are limited to single-viewpoint video, whether purely exocentric or mixed with “ego-like” views at certain time points.

We introduce Ego-Exo4D, a foundational dataset to support research on ego-exo video learning and multimodal perception. The result of a two-year effort by a consortium of 15 research institutions, Ego-Exo4D is a first-of-its-kind large-scale multimodal multiview dataset and benchmark suite. It constitutes the largest public dataset of time-synchronized first- and third-person video, captured by 740...
diverse camera wearers in 123 distinct scenes and 13 cities worldwide. For every sequence, Ego-Exo4D provides both the camera wearer’s egocentric video, as well as multiple (4-5) exocentric videos from tripods placed around the camera wearer. All views are time-synchronized and precisely localized in a metric, gravity-aligned frame of reference. The total collection has 1,286 hours of video and 5,035 instances, each spanning 1 to 42 min. of continuous capture.

Ego-Exo4D focuses on skilled single-person activities. The 740 participants perform skilled physical and/or procedural activities—dance, soccer, basketball, bouldering, music, cooking, bike repair, health care—in an unscripted manner and in natural settings (e.g., gym, soccer field, kitchens, bike shops, etc.), exhibiting a variety of skill levels from novice to expert. All video is recorded with rigorous privacy and ethics policies and formal consent of participants.

Ego-Exo4D is not only multiview, it is also multimodal. Captured with the unique open-source Aria glasses [38], all ego video is accompanied by 7-channel audio, IMU, eye gaze, both RGB and two grayscale SLAM cameras, and 3D environment point clouds. Additionally, Ego-Exo4D provides multiple new video-language resources, all time indexed: first-person narrations by the camera wearers describing their own actions; third-person play-by-play descriptions of every camera wearer action; and third-person spoken expert commentary critiquing their performance. The latter is particularly novel: performed by domain-specific experienced coaches and teachers, it focuses on how an activity is executed rather than merely what is being done, surfacing subtleties in skilled execution not perceivable by the untrained eye. To our knowledge, there is no prior video resource with such extensive and high quality multimodal data.

Alongside this data, we introduce benchmarks for foundational tasks for ego-exo video. We propose four families of tasks: 1) ego-exo relation, for relating the actions of a teacher (exo) to a learner (ego) by estimating semantic correspondences and translating viewpoints; 2) egof(-exo) recognition, for recognizing fine-grained keysteps and task structure; 3) ego(-exo) proficiency estimation, for inferring how well a person is executing a skill; and 4) ego pose, for recovering skilled 3D body and hand movements from ego-video. We provide annotations for each task—the result of more than 200,000 hours of annotator effort. To kick-start work in these new challenges, we also develop baseline models and report their results (Appendices). We are hosting the first public benchmark challenges in 2024.

In summary, Ego-Exo4D is the community’s first diverse, large-scale multimodal multiview video resource. We have open sourced all the data, annotations, camera rig protocol, and benchmarks. With this release, we aim to fuel new research in ego-exo, multimodal activity, and beyond.

2. Related work

Next we review prior work in datasets, human skill, and cross-view analysis. Section 5 will discuss related work for each benchmark task. Table 2 in Appendix 9 summarizes Ego-Exo4D’s properties vs. existing datasets.

**Egocentric datasets** There has been a surge of interest in egocentric video understanding, facilitated by recent egovideo datasets showing unscripted daily-life activity as in Ego4D [47], EPIC-Kitchens [27, 28, 163], UT Ego [78], ADL [119], and KrishnaCam [147], or procedural activities as in EGTea [81], AssistQ [172], Meccano [126], CMU-MMAC [77], and EgoProces [10]. Unlike any of the above, Ego-Exo4D focuses on multimodal ego and exo capture, and it is focused on the domain of skilled activities.

**Multiview and ego-exo datasets** Most existing multiview datasets focus on static scenes [20, 128, 151, 175, 176] and objects [133, 173], with limited (exo only) multiview human activity [26, 169]. CMU-MMAC [77] and CharadesEgo [145] are early efforts to capture both ego and exo video. CMU-MMAC [77] features 43 participants in mocap suits who cook 5 recipes in a lab kitchen. In CharadesEgo [145], 71 Mechanical Turkers record 34 hours of scripted scenarios (e.g., “type on laptop, then pick up a pillow”) from the ego and exo perspectives sequentially, yielding unsynchronized videos with non-exact activity matches. More recent ego-exo efforts focus on specific activities in one or two environments. Assembly101 [139] and H2O [76] provide time-synced ego and exo video at a lab tabletop where people assemble toy cars or manipulate handheld objects, with 53 and 4 participants, and 513 and 5 hours of footage, respectively. Homage [127] provides 30 hours of ego-exo video from 27 participants in 2 homes doing household activities like laundry.

Compared to any of the prior efforts, Ego-Exo4D offers an order of magnitude more participants, diverse locations, and hours of footage (740 participants, 123 unique scenes, 13 cities, 1,286 hours). Importantly, our focus on skilled tasks takes the participants out of the lab or home and into settings like soccer fields, dance studios, rock climbing walls, and bike repair shops. Such activities also yield a wide variety of full body poses and movements within the scene, beyond using objects at a tabletop. This variety means Ego-Exo4D augments existing 3D human body pose datasets [49, 66, 68, 80, 193]. Finally, compared to any prior ego-exo resource, Ego-Exo4D’s suite of modalities and benchmark tasks are novel and will expand the research directions the community can take for egocentric and/or exocentric video understanding.

**Human skill and video learning** Analyzing skill and action quality has received limited attention [12, 34, 35, 113, 120, 194]. Research in instructional or “how-to” videos is
facilitated by (largely exo) datasets like HowTo100M [103] and others [11, 159, 204, 207]. Challenges include grounding keysteps [10, 36, 37, 89, 103, 178, 207], procedural planning [15, 17, 22, 71, 143, 167, 196, 201], learning task structure [4, 9, 37, 107, 202, 205], and leveraging noisy narrations [89, 103, 104]. A portion of Ego-Exo4D is procedural activities, but unlike the above, it offers simultaneous ego-exo capture. The scale and diversity of our data—including its three forms of language descriptions—widen the avenues for skilled activity understanding research.

**Ego-exo cross-view modeling** There is limited prior work on ego-exo cross-view modeling, arguably due to a lack of high-quality synchronized real-world data. Prior work explores matching people between videos [5, 6, 40, 170, 179] and learning view-invariant [7, 141, 144, 182, 184, 185] or ego features [82]. Beyond the specific case of ego-exo, cross-view methods are explored for translation [130, 131, 134, 157], novel view synthesis [19, 90, 135, 137, 164, 168, 171], and aerial to ground matching [86, 132]. Ego-Exo4D provides a testbed of unprecedented size and variety for cross-view modeling. In addition, our ego-exo relation tasks (cf. Section 5) surface new challenges in novel-view synthesis with widely varying viewpoints.

### 3. Ego-Exo4D dataset

Next we introduce the dataset and its scope. Notably, the video capture was a distributed but coordinated effort performed by 12 research labs. We present the common framework, and reserve site-specific details for Appendix 10.

#### 3.1. Ego-exo camera rig

Our goal is to capture simultaneous ego and exo video, together with multiple egocentric sensing modalities. One of our contributions is to create and share a low-cost (less than $3,000), lightweight ego-exo rig with a user-friendly calibration and time sync procedure.

Our camera configuration features Aria glasses [38] for ego capture, leveraging their rich array of sensors, including an 8 MP RGB camera, two SLAM cameras, IMU, 7 microphones, and eye tracking (see Appendix 7). The ego camera is calibrated and time-synchronized with four to five (stationary) GoPros placed on tripods as the exo capture devices, allowing 3D reconstruction of the environment point clouds and the participant’s body pose. The number and placement of the exocentric cameras is determined per scenario in order to allow maximal coverage of useful viewpoints without obstructing the participants’ activity.

Our time sync and calibration design relies on a QR-code procedure to auto-sync the cameras and auto-separate the individual “takes”, meaning instances of an activity. We can do continuous recordings of up to ~60 minutes, based on the Aria battery life. See Appendix 8 for more details.

#### 3.2. Domains and environments

Ego-Exo4D focuses on skilled human activity. This is in contrast to existing ego-only efforts like Ego4D [47], which has a broad span of daily-life activities. We intentionally select the domains based on a few criteria: Will it illustrate skill and a variety of expertise? Is there visual variety to be expected across different instances? Will the ego and exo views offer complementary information? Will it present new challenges unaddressed by current datasets?

Intersecting these criteria, we arrived at two broad categories\(^1\) of skilled activity: *physical* and *procedural*, together comprising eight total domains. The physical domains are soccer, basketball, dance, bouldering, and music. They emphasize body pose and movements as well as

---

\(^1\)Note that in general physical and procedural are not mutually exclusive labels. An activity can both require physical skill and procedural steps.
interaction with objects (e.g., a ball, musical instrument). The procedural domains are cooking, bike repair, and health care. They require performing a sequence of steps to reach a goal state (e.g., a completed recipe, a repaired bike) and generally entail intricate hand-object manipulations with a variety of objects (e.g., bike repair tools; cooking utensils, appliances, and ingredients).

In total, we have 43 activities derived from the eight domains (see Appendix 9). For example, cooking is comprised of 12 recipes; soccer is comprised of 3 drills. The length of a take ranges from 8 sec to 42 min, with procedural activities like cooking having the longest sustained captures.

To achieve visual diversity in the data, multiple labs across our team (typically 3-5) captured each Ego-Exo4D domain. The data is collected in authentic settings—such as real-world bike shops, soccer pitches, or bouldering gyms—as opposed to lab environments. For example, we have videos of chefs in New York City, Vancouver, Philadelphia, Bogota, and others; soccer players in Tokyo, Chapel Hill, Hyderabad, Singapore, and Pittsburgh. See Figure 2.

3.3. Participants: expertise and diversity

We recruited 740 total participants from the local communities of 12 labs. All scenarios feature real-world experts, where the camera-wearer participant has specific credentials, training, or expertise in the skill being demonstrated. For example, among the Ego-Exo4D camera wearers are professional and college athletes; jazz, salsa, and Chinese folk dancers and instructors; competitive boulderers; professional chefs who work in industrial-scale kitchens; bike technicians who service dozens of bikes per day. Many of them have (individually) over 10 years of experience.

Experts are prioritized given they are likely to conduct activities without mistakes or distractions, providing a strong ground truth for how to approach a given task. However, we also include capture from people with varying skill levels, as well—essential for our proposed skill proficiency estimation task (Section 5). Notably, Ego-Exo4D represents human intelligence in a new way by capturing domain-specific expertise—both in the video as well as the accompanying expert commentary (see Section 4)—portraying the evolution of a skill from beginners to experts.

According to the participant surveys (Appendix 11), the camera wearers range in age from 18 to 74 years old, with 37% self-identifying as female 60% male and 3% as non-binary or preferring not to say. In total, the participants self-report more than 24 different ethnicities.

3.4. Privacy and ethics

Ego-Exo4D was collected following rigorous privacy and ethics standards. This included undergoing formal independent review processes at each institution to establish the standards for collection, management, and informed consent. Similarly, all Ego-Exo4D data collection adhered to the Project Aria Research Community Guidelines for responsible research. Since the scenarios allow for closed environments (e.g., no passerbys) nearly all video is available without de-identification. For information about each individual partners’ protocols and restrictions, please see Appendix 10. Ego-Exo4D data is gated behind a license system, which defines permitted uses, restrictions, and consequences for non-compliance.

4. Natural language descriptions

Ego-Exo4D also offers three kinds of paired natural language datasets, each time-indexed alongside the video. See Figure 3. These language annotations are not steered towards any single benchmark, but rather are a general resource that will support browsing and mining the dataset—as well as challenges in video-language learning like grounding actions and objects, self-supervised representation learning, video-conditioned language models, and skill assessment. See Appendix 12.

The first language dataset is spoken expert commentary. The goal is to reveal nuances of the skill that are not always visible to non-experts. We recruited 52 experts (distinct from the participants) to critique the recorded videos, call out strengths and weaknesses, explain how the specific behavior of the participant (e.g., hand/body pose, use of objects) affects the performance, and provide spatial markings to support their commentary. The experts are not only well-credentialed in their areas of expertise, but also have coaching or teaching experience, which facilitates clear communication. They watch the video and pause every time they have a comment, typically 7 times per minute of video. Each piece of commentary is unbounded in length, and averages 4 sentences. We provide both the transcribed speech and the raw audio (interesting for its inflection and non-word utterances), as well as the experts’ spatial drawings and numeric ratings of each participant’s skill. Videos have expert commentary by 2-5 distinct experts, offering a variety of perspectives for the same content. In total, we have 117,812 pieces of time-stamped, video-aligned commentary. These commentaries are quite novel: they focus on how the activity is executed rather than what it entails.
5. Ego-Exo4D benchmark tasks

Our second major contribution is to define the core research challenges in the domain of egocentric perception of skilled activity, particularly when ego-exo data is available for training (if not testing). To that end, we devise a suite of foundational benchmark tasks organized into four task families: relation (Sec. 5.1), recognition (Sec. 5.2), proficiency (Sec. 5.3), and ego-pose (Sec. 5.4). For each task, we provide high quality annotations and baselines that provide a starting point from which the research community can build. We will run the first formal Ego-Exo4D challenges in 2024. Due to space limits, we briefly overview each task; see the referenced Appendices for all details including baseline models and results. There are two publicly released versions of Ego-Exo4D annotations: v1 is used to train/test baselines in this paper; the larger v2 will be used for future challenge leaderboards (see Table 7 in Appendix).

5.1. Ego-exo relation

5.1.1 Ego-exo correspondence

Motivation. Establishing object-level correspondences between ego and exo viewpoints would allow AI assistants to provide visual instructions by matching third-person observations of objects from instructional videos to those in the user’s first-person view. Compared to the general correspondence problem, our setting requires tackling a number of challenges: extreme viewpoint differences, high degrees of object occlusion, and many small objects (e.g., cooking utensils and bike repair tools).

Task definition. Given a pair of synchronized ego-exo videos and a sequence of query masks of an object of interest in one of the videos, the task is to predict the corresponding mask for the same object in each synchronized frame of the other view, if it is visible. See Figure 4, left. The task can be posed with query objects in either the ego or exo video, with both directions presenting interesting challenges (e.g., high degree of occlusion in ego views, and small object size in exo views). See Appendix 13.A.1.

Related work. Related tasks are image-level sparse correspondence given query points (instead of object masks) [65] and image-level object co-segmentation [166] for jointly segmenting semantically similar objects. Our task goes beyond static object correspondence, since the interplay between human pose and object state changes during manipulation necessitate using temporal context and tracking as the query object can be highly occluded or blurry [158].
5.1.2 Ego-exo translation

**Motivation.** Our translation task entails synthesizing a target ego clip from a given exo clip. We believe this problem will drive novel research for combining recognition and object synthesis. For example, in Figure 4 (right), the approach must make effective use of the hand’s object-specific shape and appearance priors in order to synthesize the ego view of the fingertips—which are not visible in the exo clip. Furthermore, this task will stimulate advances in visual odometry, as the method must be able to infer the ego camera pose from the third-person clip. Ego-exo translation also holds strong application potential, as it may unlock the ability to generate first-person renderings of videos that were originally captured from a third-person perspective, e.g., benefitting robot perception or AR coaching.

**Task definition.** We decompose ego-exo translation into two separate tasks: ego track prediction and ego clip generation (Figure 4, right). Ego track prediction estimates the segmentation mask of an object in the unobserved ego frames given the object masks in the observed exo clip. Ego clip generation must generate the image values (i.e., RGB) within the given ground-truth ego mask by making use of the exo clip and the object masks in those frames. This decomposition effectively splits the problem into two tasks: 1) predicting the location and shape of the object in the ego clip, and 2) synthesizing its appearance given the ground-truth position. For each, we consider a variant where the pose of the ego camera with respect to the exo camera is available to use at inference time. This simplifies the problem but reduces the applicability of the method, since this information is typically not available for arbitrary third-person videos. See Appendix 13.A.2.

**Related work.** Ego-exo translation relates to cross-view image synthesis [96, 130, 157]. Within this genre, the problem of exo-to-ego generation was recently introduced for both images [93] and video [94, 97], and approached using GANs or diffusion conditioned on the input view. Our work not only formalizes this task with ample data, but its formulation also draws attention to the need for a semantic basis to new view synthesis across extreme view changes.

5.2. Ego-exo keystep recognition

This family of tasks centers around recognizing the keysteps of a procedural activity and modeling their dependencies.

5.2.1 Fine-grained keystep recognition

**Motivation.** Recognizing the step a camera wearer is performing is non-trivial: keysteps in the same activity may look similar (folding vs. smoothing the bedsheet) and may involve hand-object interactions with heavy occlusions and head motion. Models with access to multiple views during training can leverage their complementarity to account for the deficiencies of each one, by learning viewpoint invariant representations or distilling multi-view signals into a single model (e.g., human hands from ego; body pose from exo).

**Task definition.** We study ego-exo for video recognition. During training, models have access to paired ego-exo data—time-synchronized captures of the same activity from multiple known viewpoints. Each training instance has one ego view, N exo views, and a corresponding keystep label (e.g., “flip the omelette”). At test time, given only a trimmed egocentric video clip, the model must identify the keystep performed from a taxonomy of 689 keysteps across 17 procedural activities. See Figure 5, left. Importantly, all extra supervision (time-alignment, camera poses etc.) is only available at training time; inference is standard keystep recognition, but with models that benefit from cross-viewpoint training. See Appendix 13.B.1.

**Related work.** Keystep recognition has been studied in first-person [10, 126, 145, 148] or third-person [9, 100, 159, 205, 207] videos; however, limited work considers both views together. Prior work considers cross-view learning with unpaired videos [7, 82, 182] and view-invariant feature learning on paired videos [144]. In contrast, we ex-
explore keystep recognition in large-scale, procedural activities with fully synchronized training videos.

5.2.2 Energy-efficient multimodal keystep recognition

Motivation. Current activity detection models assume access to densely sampled clips from the full video and ample computational resources to process them. These assumptions are incompatible with real-world devices (e.g., mobile phones, AR glasses) where the camera is not always on and the compute budget is limited by battery life. This task focuses on building energy-efficient video models to pave the way for feasibility on real-world hardware.

Task definition. We formulate the problem as an online action detection task, with a given energy budget. See Figure 5, right. Given a stream of audio, IMU, and RGB video data, a model must identify the keystep being performed at each frame, as well as decide which sensor(s) to use for subsequent time-steps. This task will inspire models that are strategic about which modality to deploy when. Energy consumption is the sum of sensor energy (operating the camera/audio/IMU sensors), model inference costs, and memory transfer costs, and must be within 20mW to reflect real-world device power constraints. See Appendix 13.B.2.

Related work. Prior work on efficient models considers light-weight architectures [41, 56, 101, 155, 165, 195], efficient input processing [43, 44, 73, 102, 156], or inference optimizations [39, 58, 121, 174, 206]. In all cases, they optimize computation (FLOPs), parameter count, or prediction throughput (FPS), which in isolation are insufficient to characterize running on real-world devices. To address this, we propose the first benchmark for energy-efficient video recognition that is tied to real-world, on-device constraints, and measures total power consumed.

5.2.3 Procedure understanding

Motivation. Automatically understanding the structure of a procedure from video (inferring keystep ordering, preconditions, etc.) would allow assisting AR users in a task or informing robots that learn from human demonstrations.

Task definition. In our procedure understanding task, given a video segment $s_t$ and its previous video segment history, models have to 1) determine previous keysteps (to be performed before $s_t$); infer if $s_t$ is 2) optional or 3) a procedural mistake; 4) predict missing keysteps (should have been performed before $s_t$ but were not); and 5) next keysteps (for which dependencies are satisfied). The task offers two versions of weak supervision: instance-level: segments and their keystep labels are available for train/test; and procedure-level: only unlabeled segments and procedure-specific keystep names are given for train/test. See Figure 5 (center) and Appendix 13.B.3.

Related work. Prior work focusing on procedural understanding learns an explicit graph [62, 150, 177] as ground truth or uses a task graph for representation learning [9, 107, 202] and short-term step understanding [9, 36, 202]. Other work [32, 139] studies mistake detection in a supervised setting. We are the first to propose procedural understanding to evaluate the long-term structure of the task in a weakly-supervised setting.

5.3. Ego-exo proficiency estimation

Motivation. Going beyond recognizing what a person is doing, this task aims to infer the user’s skill level. Such an ability could lead to novel coaching tools that let people learn new skills more effectively, or new ways to evaluate human performance in domains like sports or music.

Task definition. We consider two variants: 1) demonstrator and (2) demonstration proficiency estimation. Both tasks consider one egocentric and (optionally) $M$ exocentric videos synchronized in time as their inputs. Demonstrator proficiency is formulated as a video classification task, where the model has to output one of four labels (novice, early, intermediate, or late expert). Demonstration proficiency is formulated as a temporal action localization task where given an untrimmed video, the model must output a list of tuples, each containing a timestamp, a proficiency category (i.e., good execution or needs improvement), and its probability. Note that parts of the video that do not reveal the participant’s skill are left unlabeled. See Figure 6 and Appendix 13.C.

Related work. Prior work uses egocentric [12, 35] or exocentric [60, 114, 115] views for proficiency estimation in sports [12, 115, 120], health [60, 92, 191, 208], and others [35, 186]. We propose the first multi-view egocentric and exocentric proficiency estimation benchmark. Unlike prior work, our benchmark spans diverse, day-to-day physical and procedural scenarios and includes temporally localized annotations of (in)correct executions.

5.4. Ego pose

This family of tasks is motivated by recovering the skilled body movements of participants, even in the extreme setting of monocular ego-video input in dynamic environments.
Motivation. Estimating the physical state of a person’s body—the 3D positions of the arms, legs, hands—from the ego view is essential for wearable AI systems that can support human activity. Challenges include subtle and flexible movements, frequent occlusion, and body parts out of view. Task definition. For both the body and hand pose (“ego pose”) estimation tasks, the input is an ego video. The output is a set of 3D joint positions of the camera wearer’s body and hands for each time step, parameterized as 17 3D body joint positions and 21 3D joint positions per hand, following the MS COCO convention [87]. To our knowledge, Ego-Exo4D offers the largest manual ground truth (GT) egocentric body and hand pose annotations to date. And, in total, it offers ~14M frames of 3D GT and pseudo-GT combined. See Figure 7 and Appendix 13.D. Related work. Limited prior work explores 3D body pose from a wearable camera. Some methods assume no body visibility [63, 80, 98, 187, 188], while others assume partial observability by modifying cameras to capture the body [3, 57, 136, 161, 181]. Our dataset can be used for both paradigms. Existing hand pose datasets use constrained environments [106, 146] with simple hand motion [50, 76, 109], whereas we include diverse real-world scenarios, e.g., with expert musicians and bike mechanics. 6. Conclusions

Ego-Exo4D provides a dataset of unprecedented scale and realism for ego-exo video learning. It offers a unique window into skilled human activity from 8 compelling domains by hundreds of real-world experts around the globe. Together with the proposed benchmarks, we hope that this new open source resource will set the stage for substantial new research for the years to come. Though we are motivated by skill learning, Ego-Exo4D is poised for even broader influence, beyond the proposed benchmarks. Whereas existing datasets lack activity modeling in real-world 3D contexts (e.g., restricted to mocap suits and/or lab settings). Ego-Exo4D is a resource for general 3D vision—such as environment reconstruction, camera re-localization, audio-visual mapping, and many others. Similarly, our novel video-language resources will offer many opportunities for grounding of actions and objects, multi-modal representation learning, and language generation. Finally, though our tasks prioritize perception from the “ego-only” perspective, the ego component of our data ensures its utility for the more traditional exo viewpoint too, e.g., for activity recognition and body pose estimation.

Contribution statement

This project is the result of a large collaboration between many institutions over the last two years. Initial authors represent the leadership team of the project. Kristen Grauman initiated the project, served as the technical lead, initiated the recognition and proficiency benchmarks and expert commentary, and coordinated their working groups. Andrew Westbury served as the program manager and operations lead for all aspects of the project. Lorenzo Torresani led development of the capture domains, initiated the relation and ego-pose benchmarks, and coordinated their working groups. Kris Kitani led development of the multi-camera rig and supported the Ego-Exo4D engineering team on all aspects of the data annotation and organization. Jitendra Malik served as a scientific advisor. Authors with stars (*) were key drivers of implementation, collection, and/or annotation development throughout the project. Authors with daggers (†) are faculty and senior researcher PIs for the project. The Appendices detail the contributions of individual authors for the various benchmarks, data collection, and annotation pipelines.

Acknowledgements

We gratefully acknowledge the following colleagues for valuable discussions and support of our project: Vittorio Caggiano, Ilé Danza, Ahmad Darkhalil, Zona de Bloque, Alex Dinh, Rene Martinez Doehner, Ivan Cruz, Matt Feiszli, Vance Feutz, Kelly Forbes, Rohit Girdhar, Pierre Gleize, Andrés Hernández, Shun Iwase, Bolin Lai, Vivian Lee, Brighid Meredith, Ashley Massie, Natalia Neverova, Manohar Paluri, Joelle Pineau, Artiom Sanakoyeu, Paresh Shenoy, Jiary Sh, Jiasheng Shi, Gaurav Shrivastava, Mitesh Singh, Manasi Swaminathan, Arjang Talattof, Ali Thabet, Laurens van der Maaten, Andrea Vedaldi, and Tobby Zhu. We also sincerely thank the 52 experts who contributed to the expert commentary for their expertise and support; they are listed individually in Appendix 12. Thank you to the Common Visual Data Foundation (CVDF) for hosting the Ego-Exo4D dataset. Finally, thank you to the 740 participants who contributed to this dataset and shared their skills in video.

Univ. of Bristol is supported in part by EPSRC UMPIRE (EP/T004991/1) and EPSRC PG Visual AI (EP/T028572/1). Z. Zhu is supported by UoB-CSC Scholarship. Univ. of Catania is supported in part by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006. Simon Fraser Univ. is supported in part by the Canada Research Chairs Program (CRC-2019-00298) and NSERC Discovery (2019-06489). Georgia Tech is supported in part by NSF award CNS-2308994. IU is supported in part by NSF DRL-2112635 (AI Institute for Engaged Learning). UT Austin is supported in part by the IFML NSF AI Institute.
References


[16] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvp: Video-text pre-training for temporal localization. In European Conference on Computer Vision, 2022. 4


[47] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Iljia Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhart Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph


[55] Mark Horowitz. 1.1 computing’s energy problem (and what we can do about it). In 2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC), 2014. 46.


[70] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar,


[81] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In ECCV, 2018. 3, 7


[86] TY Lin, Y Cui, S Belongie, and J Hays. Learning deep representations for ground-to-aerial geolocalization. In CVPR, 2015. 4


[98] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Ilya Loshchilov and Frank Hutter. Decoupled weight de-


[100] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations. 2022. 7


[105] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfriend, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. PAMI, 2019. 2


[110] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representa-


[114] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events, 2017. 8


[118] Ethan Perez, Florian Strub, Harm De Vries, Vincent Du-


[122] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukávs, Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hamernik, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The kaldí speech recognition toolkit. 2011. 47

[123] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou,


[148] Yale Song, Eugene Byrne, Tushar Nagarajan, Huuyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step:
Advances in Neural Information Processing Systems (ECCV)


