

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

Kristen Grauman^{1,2}, Andrew Westbury¹, Lorenzo Torresani¹, Kris Kitani^{1,3}, Jitendra Malik^{1,4}, Triantafyllos Afouras^{*1}, Kumar Ashutosh^{*1,2}, Vijay Baiyya^{*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 Forigua^{*12}, Abrham 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}, Effrosyni Mavroudi^{*1}, Tushar Nagarajan^{*1}, Francesco Ragusa^{*15}, Santhosh Ramakrishnan^{*2}, Luigi Seminara^{*15}, Arjun Somayazulu^{*2}, Yale Song^{*1}, Shan Su^{*16}, Zihui Xue^{*1,2}, Edward Zhang^{*16}, Jin Xu Zhang^{*16}, Angela Castillo¹², Changan Chen², Xinzhu Fu¹¹, Ryosuke Furuta¹⁷, Cristina González¹², Prince Gupta⁵, Jiabo Hu¹⁸, Yifei Huang¹⁷, Yiming Huang¹⁶, Leslie Khoo¹⁹, Anush Kumar¹⁰, Robert Kuo¹⁸, Sach Lakhavani⁵, Miao Liu¹⁸, Mi Luo², Zhengyi Luo³, Brigid Meredith¹⁸, Austin Miller¹⁸, Oluwatumininu Oguntola¹⁴, Xiaqing Pan⁵, Penny Peng¹⁸, Shraman Pramanick²⁰, Merey Ramazanova²¹, Fiona Ryan²², Wei Shan¹⁴, Kiran Somasundaram⁵, Chenan Song¹¹, Audrey Southerland²², Masatoshi Tateno¹⁷, Huiyu Wang¹, Yuchen Wang¹⁹, Takuma Yagi¹⁷, Mingfei Yan⁵, Xitong Yang¹, Zecheng Yu¹⁷, Shengxin Cindy Zha¹⁸, Chen Zhao²¹, Ziwei Zhao¹⁹, Zhifan Zhu⁶, Jeff Zhuo¹⁴, Pablo Arbeláez^{†12}, Gedas Bertasius^{†14}, David Crandall^{†19}, Dima Damen^{†6}, Jakob Engel^{†5}, Giovanni Maria Farinella^{†15}, Antonino Furnari^{†15}, Bernard Ghanem^{†21}, Judy Hoffman^{†22}, C. V. Jawahar^{†7}, Richard Newcombe^{†5}, Hyun Soo Park^{†10}, James M. Rehg^{†8}, Yoichi Sato^{†17}, Manolis Savva^{†13}, Jianbo Shi^{†16}, Mike Zheng Shou^{†11}, and Michael Wray^{†6}

¹FAIR, 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, ¹²Universidad de los Andes, ¹³Simon Fraser University,

¹⁴University of North Carolina, Chapel Hill, ¹⁵University of Catania, ¹⁶University of Pennsylvania, ¹⁷University of Tokyo,

¹⁸Meta, ¹⁹Indiana University, ²⁰Johns Hopkins University, ²¹King Abdullah University of Science and Technology, ²²Georgia Tech

Abstract

We present Ego-Exo4D, a diverse, large-scale multi-modal 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). More than 800 participants from 13 cities worldwide performed these activities in 131 different natural scene contexts, yielding long-form captures from 1 to 42 minutes each and 1,422 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 hu-

man 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 will be 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?

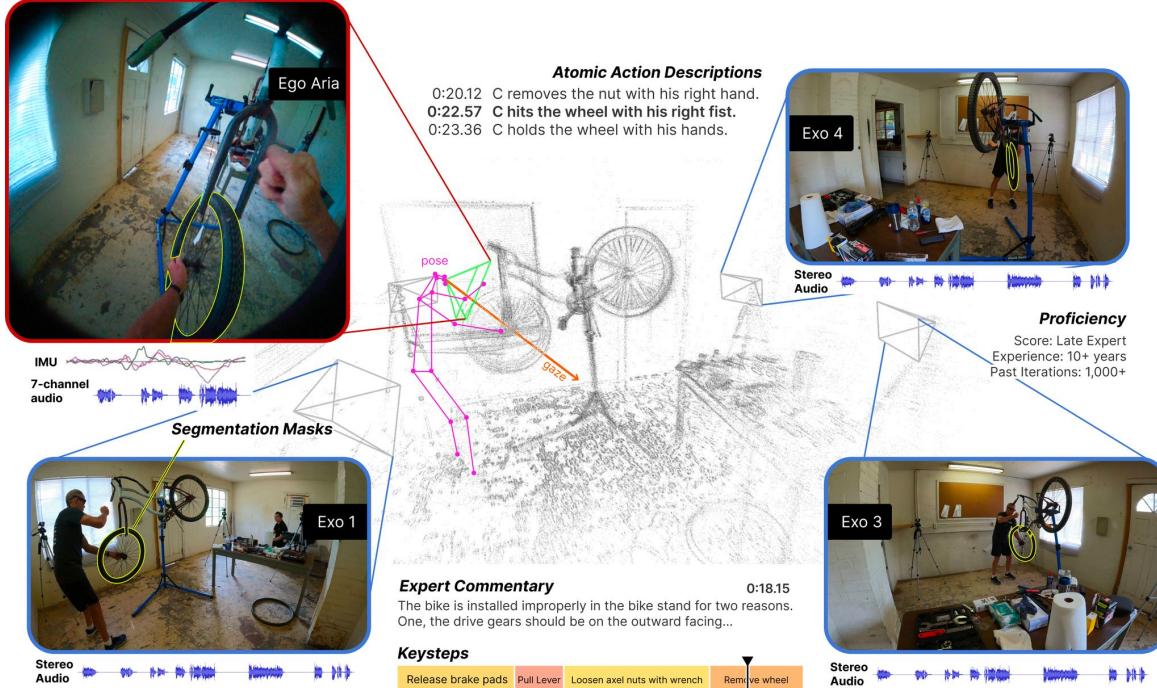


Figure 1. Ego-Exo4D offers egocentric video alongside multiple time-synchronized exocentric video streams for an array of skilled human activities—1,422 hours of ego and exo video in total. The data is both multiview and multimodal, and it is extensively annotated with language, 3D body and hand pose, keysteps, procedural dependencies, and proficiency ratings in support of our proposed benchmark tasks.

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) [39, 104], 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 [73, 74, 123, 135, 141], 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 [26, 44] or exo [45, 64, 101, 145] view, leaving the ability to move fluidly between the first- and third-person perspectives out of reach. Instructional video datasets [99, 155, 198, 201] 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 multiview dataset and benchmark suite. It constitutes the largest public dataset of time-synchronized first- and third- person video, captured by 839

diverse camera wearers in 131 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 1,422 hours of video and 5,625 instances, each spanning 1 to 42 min. of continuous capture.

Ego-Exo4D focuses on skilled single-person activities. The 839 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 [35], 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 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) *ego(-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 the skilled 3D body and hand movements of experts from ego-video. We provide annotations for each task—the result of more than 200,000 hours of annotator effort. To kickstart work in these new challenges, we also develop baseline models and report their results. We plan to host a first public benchmark challenge in June 2024.

In summary, Ego-Exo4D is the community’s first diverse, large-scale multimodal multiview video resource. We will open source all the data, annotations, camera rig protocol, benchmark tasks, and baseline code. 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 of the benchmark tasks in turn.

Egocentric datasets There has been a surge of interest in egocentric video understanding, facilitated by recent ego-video datasets showing unscripted daily-life activity as in Ego4D [44], EPIC-Kitchens [25, 26, 159], UT Ego [75], ADL [115], and KrishnaCam [143], or procedural activities as in EGTea [78], AssistQ [168], Meccano [122], CMU-MMAC [74], and EgoProcel [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 [19, 124, 147, 170, 171] and objects [129, 169], with limited (exo only) multiview human activity [24, 165]. CMU-MMAC [74] and CharadesEgo [141] are early efforts to capture both ego and exo video. CMU-MMAC [74] features 43 participants in mocap suits who cook 5 recipes in a lab kitchen. In CharadesEgo [141], 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 [135] and H2O [73] 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 [123] 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 (839 participants, 131 unique scenes, 13 cities, 1,422 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 also complements existing 3D human body pose datasets [46, 63, 65, 77, 188]. 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, 31, 32, 109, 116, 189]. Research in instructional or “how-to” videos is facilitated by (largely exo) datasets like HowTo100M [99]

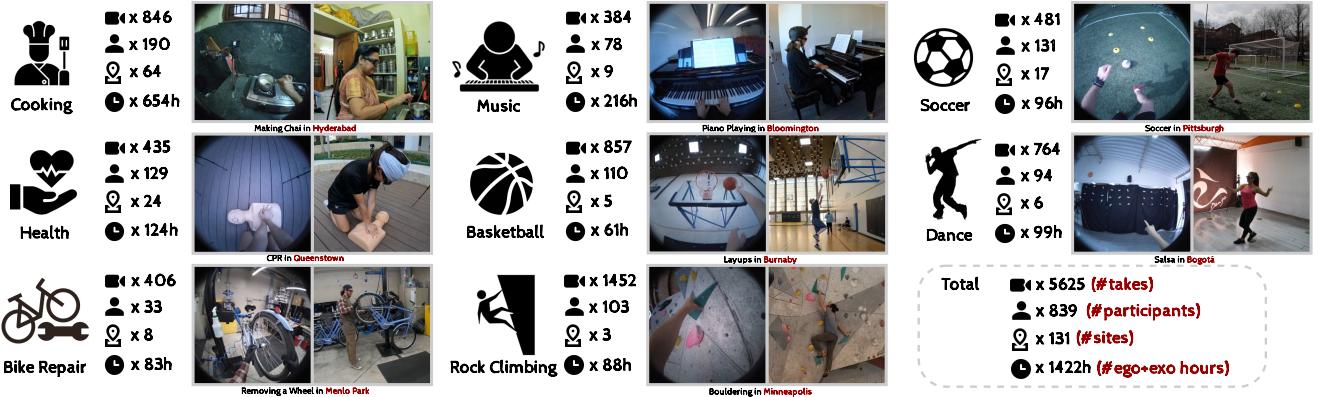


Figure 2. Ego-Exo4D captures skilled activity from 8 domains, in a wide variety of 131 scenes in 13 different cities in Japan, Colombia, Canada, India, Singapore, and seven US states. Each domain is captured at multiple sites—at least 3 and as many as 64 unique locations. In total the dataset offers 1,422 hours of ego+exo video comprised of 5,625 takes from 839 camera wearers. An average take is 3 minutes.

and others [11, 155, 198, 201]. Challenges include grounding keysteps [10, 33, 34, 86, 99, 100, 173, 201], procedural planning [14, 16, 20, 68, 139, 163, 191, 196], learning task structure [4, 9, 34, 103, 197, 199], and leveraging noisy narrations [86, 99, 100]. 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, 37, 166, 174] and learning view-invariant [7, 137, 140, 177, 179, 180] or ego features [79]. Beyond the specific case of ego-exo, cross-view methods are explored for translation [126, 127, 130, 153], novel view synthesis [18, 87, 131, 133, 160, 164, 167], and aerial to ground matching [83, 128]. 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 [35] 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 \sim 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 [44], 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 categories¹ 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 interaction with objects (e.g., a ball, musical instrument). The procedural domains are cooking, bike repair, and health

¹Note that in general physical and procedural are not mutually exclusive labels. An activity can both require physical skill and procedural steps.

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 14 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

We recruited 839 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 re-



Figure 3. Ego-Exo4D offers three paired language corpora. Word cloud is from expert commentary which critique the performance.

sponsible 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 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. All videos have expert commentary by 2-5 distinct experts, offering a variety of perspectives for the same content. In total, we have 41,087 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, capturing subtle differences in skilled execution. We be-

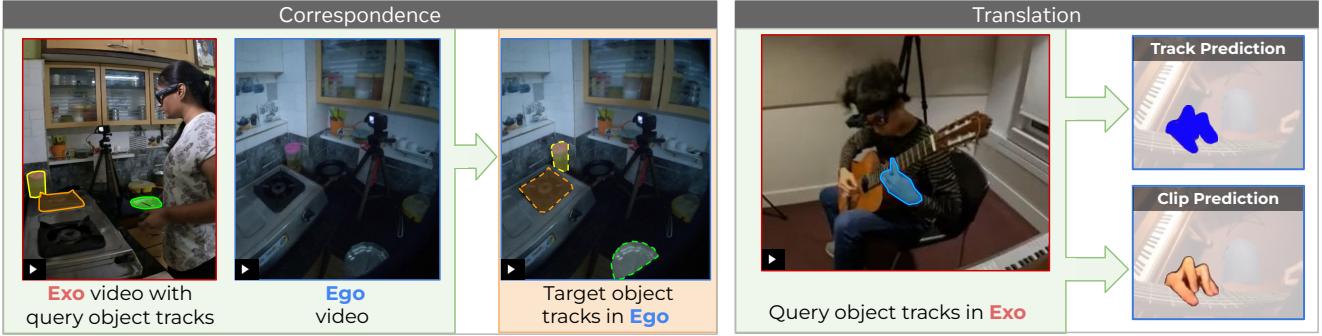


Figure 4. The ego-exo relation family consists of the tasks of correspondence (left) and translation (right).

lieve this can unlock new fundamental problems (e.g., proficiency estimation below) and disruptive future applications (e.g., AI coaching).

The second language dataset consists of *narrate-and-act descriptions* provided by the participants themselves. They are in the style of a tutorial or how-to video, where the participant explains what they are doing and why. Unlike the third-party expert commentary above, these are first-person reflections on the activity given by the people doing them. These narrations are available for about 10% of all takes in the dataset, since we wanted participants to execute the tasks without pausing for the bulk of the recordings.

The third language dataset consists of *atomic action descriptions*. Whereas the commentary and narrate-and-act language reveals spoken opinions and reasons for the actions (the “why and how”), this stream of text is specifically about the “what”. Inspired by Ego4D’s narrations [44], these are short statements written by third-party annotators, timestamped for every atomic action performed by the participant. They are written by two independent (non-domain expert) annotators for all video in the dataset, for a total of 150K sentences. This data is valuable for mining for taxonomies of objects and actions in the data, indexing the videos with keywords for exploring the dataset, and for future research in video-language learning, as has been quite successful for the Ego4D narrations [8, 82, 119].

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 aim to run the first formal Ego-Exo4D challenge in 2024. Due to space limits, here we briefly overview

each task; see the referenced Appendices for all details including baseline models and results.

5.1. Ego-exo relation

Our ego-exo *relation* tasks deal with relating the video content across the extreme ego-exo viewpoint changes. They take the form of object-level matching (correspondence) and synthesis of one view from the other (translation).

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) [62] and image-level object co-segmentation [162] 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 [154].

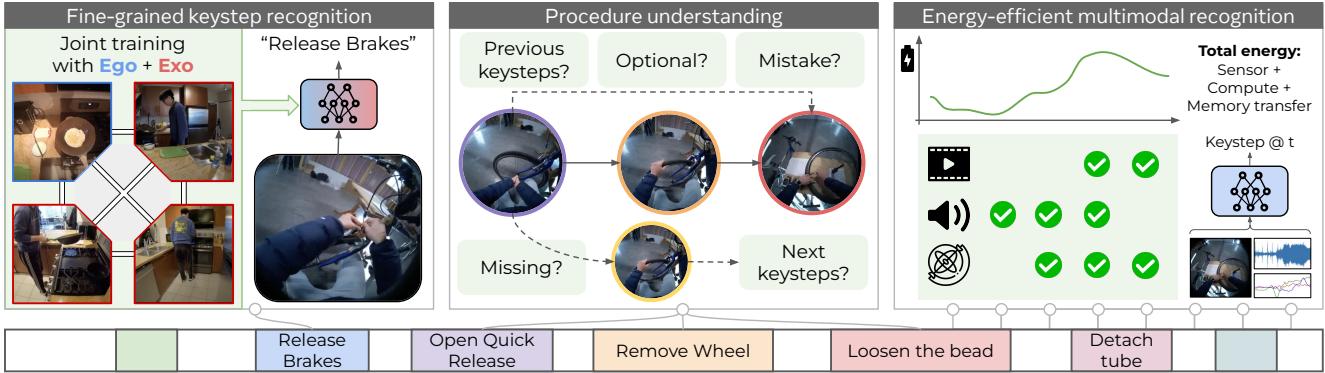


Figure 5. **Ego-exo keystep recognition.** This family of tasks consists of fine-grained recognition (left, Section 5.2.1), procedure understanding (center, Section 5.2.3) and energy-efficient multimodal recognition (right, Section 5.2.2).

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 [93, 126, 153]. Within this genre, the problem of exo-to-ego generation was recently introduced for both images [90] and video [91], and approached using GANs 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, unknown 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, 122, 141, 144] or third-person [96, 155, 199, 201] videos; however, limited work considers both views together. Prior work considers cross-view learning with unpaired videos [7, 79, 177] and view-invariant fea-

ture learning on paired videos [140]. In contrast, we 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 [38, 53, 97, 151, 161, 190], efficient input processing [40, 41, 70, 98, 152], or inference optimizations [36, 55, 117, 200]. 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 measure 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 version 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.

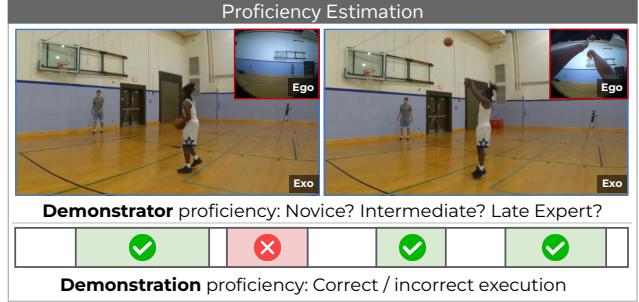


Figure 6. Demonstrator and demonstration proficiency estimation.

Related work. Prior work focusing on procedural understanding learns an explicit graph [59, 146, 172] as ground truth or uses a task graph for representation learning [9, 103, 197] and short-term step understanding [9, 33, 197]. Other work [29, 135] 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 has to 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.

Related work. Prior work uses egocentric [12, 32] or exocentric [57, 110, 111] views for proficiency estimation in sports [12, 111, 116], health [57, 89, 186, 202], and others [32, 181]. 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.

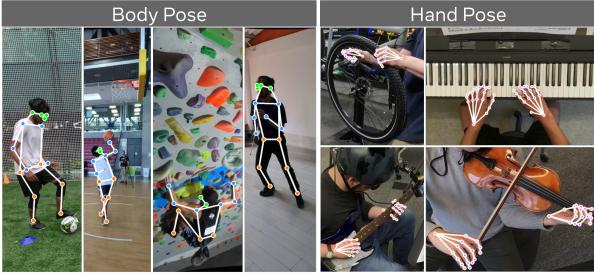


Figure 7. Hand and body keypoints for ego-pose estimation

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 [84]. See Figure 7.

Related work. Limited prior work explores 3D body pose from a wearable camera. Some methods assume no body visibility [60, 77, 94, 182, 183], while others assume partial observability by modifying cameras to capture the body [3, 54, 132, 157, 176]. Our dataset can be used for both paradigms. Existing hand pose datasets use constrained environments [102, 142] with simple hand motion [47, 73, 105], 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 relocalization, 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 exo 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: Ilé Danza, Ahmad Darkhalil, Zona de Bloque, Ivan Cruz, Matt Feiszli, Vance Feutz, Rohit Girdhar, Pierre Gleize, Andrés Hernández, Vivian Lee, Ashley Massie, Natalia Neverova, Joelle Pineau, Jiaray Shi, Jiasheng Shi, Mitesh Singh, Manasi Swaminathan, 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 839 participants who contributed to this dataset and shared their skills in video.

The University of Bristol is supported in part by EPSRC UMPIRE (EP/T004991/1) and EPSRC PG Visual AI (EP/T028572/1). Zhifan Zhu is supported by UoB-CSC Scholarship.

References

- [1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed El-hayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *Proceedings of the*

- IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1001–1010, 2023. 57
- [2] Michael Abrash. Creating the future: Augmented reality, the next human-machine interface. In *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021. 45
- [3] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. Mecap: Whole-body digitization for low-cost vr/ar headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 453–462, 2019. 9
- [4] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 4
- [5] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV*, 2016. 4
- [6] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE transactions on pattern analysis and machine intelligence*, 41(6), 2018. 4
- [7] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171, 2018. 4, 7
- [8] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6
- [9] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. In *NeurIPS*, 2023. 4, 8, 28
- [10] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 7
- [11] Yizhak Ben-Shabat, Xin Yu, Fatemehsadat Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. 2020. 4
- [12] Gedas Bertasius, Hyun Soo Park, Stella Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *ICCV*, 2017. 3, 8
- [13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 38, 42, 43, 46, 53, 54
- [14] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 4
- [15] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010. 34
- [16] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtpt: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, 2022. 4
- [17] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. *CV4Metaverse workshop, International Conference on Computer Vision*, 2023. 55, 56
- [18] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 4
- [19] Angel Chang, Angela Dai, Tom Funkhouser, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgbd data in indoor environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017. MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf. 3
- [20] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 334–350. Springer, 2020. 4
- [21] Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, and Vivienne Sze. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019. 45
- [22] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 34, 35
- [23] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/openmmlab/mmpose>, 2020. 55
- [24] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *WACV*, 2021. 3
- [25] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3, 26
- [26] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *IJCV*, 2021. 2, 3
- [27] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 269–284. Springer, 2016. 45, 50

- [28] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023. 45
- [29] Guodong Ding, Fadime Sener, Shugao Ma, and Angela Yao. Every mistake counts in assembly. *arXiv preprint arXiv:2307.16453*, 2023. 8
- [30] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 38
- [31] H Doughty, D Damen, and W Mayol-Cuevas. Who’s better? who’s best? pairwise deep ranking for skill determination. In *CVPR*, 2018. 3
- [32] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. 2019. 3, 8
- [33] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *ECCV*, pages 319–335. Springer, 2022. 4, 8
- [34] Ehsan Elhamifar and Dat Huynh. Self-supervised multi-task procedure learning from instructional videos. In *European Conference on Computer Vision*, pages 557–573. Springer, 2020. 4
- [35] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gumno, Andrew Turner, Arjang Tattnof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulou, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Baltanás, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project Aria: A new tool for egocentric multi-modal AI research, 2023. 3, 4, 2
- [36] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019. 8
- [37] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *CVPR*, 2017. 4
- [38] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 8, 44, 45, 46, 47
- [39] John H. Flavell, Eleanor R. Flavell, Frances L. Green, and Sharon A. Wilcox. The development of three spatial perspective-taking rules. *Child Development*, 1981. 2
- [40] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 8
- [41] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8
- [42] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 53
- [43] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The “ something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 40
- [44] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragnani, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Leslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2, 3, 4, 6, 28, 32, 42
- [45] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A

- video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2
- [46] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human positioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, 2021. 3
- [47] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnorate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 9
- [48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 35, 57
- [49] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 42, 43
- [50] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 55
- [51] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 38
- [52] 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. 45
- [53] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 8
- [54] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 98–111, 2020. 9
- [55] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 8
- [56] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 25
- [57] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 214–221, 2018. 8
- [58] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 38
- [59] Yunseok Jang, Sungryull Sohn, Lajanugen Logeswaran, Tiange Luo, Moontae Lee, and Honglak Lee. Multimodal subtask graph generation from instructional videos. *arXiv preprint arXiv:2302.08672*, 2023. 8
- [60] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *CVPR*, 2017. 9
- [61] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*, pages 443–460. Springer, 2022. 55, 56
- [62] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 6, 34
- [63] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3
- [64] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 40, 42
- [65] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. In *ICCV*, 2023. 3
- [66] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 35
- [67] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 32
- [68] Dohwan Ko, Joomyoung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5016–5025, 2022. 4
- [69] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 57
- [70] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 8
- [71] Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon. Efficient intersection of three quadrics and applications in computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1799–1808, 2016. 3

- [72] Iljung Kwak, Jian-Zhong Guo, Adam Hantman, David Kriegman, and Kristin Branson. Detecting the starting frame of actions in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 489–497, 2020. 53
- [73] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 2, 3, 9
- [74] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the carnegie mellon university multimodal activity (cmu-mmact) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, 2009. 2, 3
- [75] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 3
- [76] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 57
- [77] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 3, 9, 55, 56
- [78] 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
- [79] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 4, 7, 20, 42, 43
- [80] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyu Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22932–22941, 2023. 46, 47
- [81] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 57
- [82] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 2022. 6, 25, 49
- [83] TY Lin, Y Cui, S Belongie, and J Hays. Learning deep representations for ground-to-aerial geolocation. In *CVPR*, 2015. 4
- [84] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 9
- [85] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 57
- [86] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 4, 28
- [87] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 4
- [88] Chiao Liu, Lyle Bainbridge, Andrew Berkovich, Song Chen, Wei Gao, Tsung-Hsun Tsai, Kazuya Mori, Rimon Ikeno, Masayuki Uno, Toshiyuki Isozaki, et al. A $4.6\ \mu\text{m}$, 512×512 , ultra-low power stacked digital pixel sensor with triple quantization and 127db dynamic range. In *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020. 45
- [89] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9522–9531, 2021. 8
- [90] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1843–1847. IEEE, 2020. 7
- [91] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021. 7
- [92] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 47
- [93] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7
- [94] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *Advances in Neural Information Processing Systems*, 2021. 9, 55, 56
- [95] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 55
- [96] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations. 2022. 7

- [97] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 8
- [98] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *ECCV 2020*, 2020. 8
- [99] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2, 3, 4, 28
- [100] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 4, 28
- [101] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *PAMI*, 2019. 2
- [102] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 9
- [103] Medhini Narasimhan, Licheng Yu, Sean Bell, Ning Zhang, and Trevor Darrell. Learning and verification of task structure in instructional videos. *arXiv preprint arXiv:2303.13519*, 2023. 4, 8
- [104] Nora Newcombe. The development of spatial perspective taking. *Advances in child development and behavior*, 1989. 2
- [105] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12999–13008, 2023. 9
- [106] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 42, 43
- [107] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020. 25
- [108] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 57
- [109] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, 2019. 3
- [110] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events, 2017. 8
- [111] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 8
- [112] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 38
- [113] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 34, 38
- [114] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 38
- [115] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 3
- [116] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *ECCV*, 2014. 3, 8
- [117] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018. 8
- [118] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The kaldi speech recognition toolkit. 2011. 46
- [119] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 6, 25, 42, 43
- [120] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 38
- [121] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavy, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 26
- [122] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *WACV*, 2021. 3, 7
- [123] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J.C. Niebles. Home action genome: Contrastive compositional action understanding. In *CVPR*, 2021. 2, 3
- [124] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M

- Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. [3](#)
- [125] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. [38](#)
- [126] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [4, 7](#)
- [127] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 2019. [4](#)
- [128] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *ICCV*, 2019. [4](#)
- [129] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. [3](#)
- [130] Bin Ren, Hao Tang, and Nicu Sebe. Cascaded cross mlpmixer gans for cross-view image translation. *arXiv preprint arXiv:2110.10183*, 2021. [4](#)
- [131] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022. [4](#)
- [132] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. [9](#)
- [133] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021. [4](#)
- [134] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. [57](#)
- [135] Fadime Sener, Dipyadip Chatterjee, Daniel Shelepor, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. [2, 3, 8](#)
- [136] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 486–487. IEEE, 2017. [42](#)
- [137] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018. [4](#)
- [138] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022. [34, 35](#)
- [139] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. [4](#)
- [140] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018. [4, 8, 42](#)
- [141] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. [2, 3, 7](#)
- [142] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. [9](#)
- [143] Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A. Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *WACV*, 2016. [3](#)
- [144] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *NeurIPS*, 2023. [7](#)
- [145] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012. [2, 40](#)
- [146] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *ICCV*, pages 4669–4677, 2015. [8](#)
- [147] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. [3](#)
- [148] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction*

- with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 38
- [149] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. How to evaluate deep neural network processors: Tops/w (alone) considered harmful. *IEEE Solid-State Circuits Magazine*, 2020. 45
- [150] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that neRF needs? In *The Eleventh International Conference on Learning Representations*, 2023. 38
- [151] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019. 8
- [152] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *NeurIPS*, 2023. 8
- [153] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019. 4, 7
- [154] Hao Tang, Kevin Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 2023. 6
- [155] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The coin dataset and performance evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 4, 7
- [156] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 55
- [157] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 9
- [158] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. 2022. 42
- [159] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023. 3
- [160] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. *arXiv preprint arXiv:2303.17598*, 2023. 4
- [161] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Mobileone: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7907–7917, 2023. 8
- [162] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR 2011*, pages 2217–2224. IEEE, 2011. 6
- [163] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. Pdpp: Projected diffusion for procedure planning in instructional videos. *arXiv preprint arXiv:2303.14676*, 2023. 4
- [164] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 4
- [165] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding (CVIU)*, 2006. 3
- [166] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer’s location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3446–3455, 2021. 4
- [167] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 4
- [168] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Asstsq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision*, 2022. 3
- [169] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2015. 3
- [170] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *CVPR*. IEEE, 2018. Gibson license is available at http://svl.stanford.edu/gibson2/assets/GDS_agreement.pdf. 3
- [171] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 3
- [172] Frank F Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. A benchmark for structured procedural knowledge extraction from cooking videos. *arXiv preprint arXiv:2005.00706*, 2020. 8
- [173] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 4
- [174] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *ECCV*, 2018. 4

- [175] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. [48](#)
- [176] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. [9](#)
- [177] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *NeurIPS*, 2023. [4](#), [7](#)
- [178] Lita Yang, Robert M Radway, Yu-Hsin Chen, Tony F Wu, Huichu Liu, Elnaz Ansari, Vikas Chandra, Subhasish Mitra, and Edith Beigné. Three-dimensional stacked neural network accelerator architectures for ar/vr applications. *IEEE Micro*, 2022. [45](#)
- [179] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *ACM MM*, 2019. [4](#)
- [180] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First- and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [4](#)
- [181] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7919–7928, 2021. [8](#)
- [182] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. [9](#)
- [183] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. [9](#)
- [184] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *arXiv preprint arXiv:2202.07925*, 2022. [20](#)
- [185] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510, 2022. [53](#), [54](#), [55](#)
- [186] Qiang Zhang and Baoxin Li. Relative hidden markov models for evaluating motion skill. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. [8](#)
- [187] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [38](#)
- [188] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. [3](#)
- [189] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *CVPR*, 2023. [3](#)
- [190] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. [8](#)
- [191] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2938–2948, 2022. [4](#)
- [192] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20438–20447, 2022. [57](#)
- [193] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, 2022. [48](#)
- [194] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. [25](#), [45](#), [46](#), [47](#)
- [195] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1620, 2023. [57](#)
- [196] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueteng Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. *arXiv preprint arXiv:2303.17839*, 2023. [4](#)
- [197] Honglu Zhou, Roberto Martin-Martin, Mubbashir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. [4](#), [8](#)
- [198] L. Zhou, N. Louis, and J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*, 2018. [2](#), [4](#)
- [199] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. [4](#), [7](#)
- [200] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017. [8](#)
- [201] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. [2](#), [4](#), [7](#)
- [202] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, and Irfan A. Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *CoRR*, abs/1702.07772, 2017. [8](#)

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

Supplementary Material

Appendix

Table of Contents

Appendices	1
7. Aria Glasses	2
7.A Device and Sensors	2
7.B Machine Perception Services (MPS) .	2
7.C Processing Summary	4
7.D Tools and Ecosystem	4
8. Camera Rig and Data Processing	5
8.A Hardware	5
8.B Time Sync	5
8.C Take Separation	5
8.D Recording Procedure	5
8.E Extensions	6
9. Scenarios	7
10 Data Collection	11
10.A Carnegie Mellon University	11
10.B FAIR, Meta	12
10.C Georgia Tech	12
10.D IIIT-Hyderabad	13
10.E Indiana University	14
10.F National University Singapore	15
10.G Simon Fraser University	15
10.H Universidad de los Andes	16
10.I University of Minnesota	17
10.J University of North Carolina	17
10.K University of Pennsylvania	18
10.L University of Tokyo	19
11 Participants	22
11.A Demographics	22
11.B Participant Surveys	23
12 Language Descriptions	25
12.A Expert Commentary	25
12.B Narrate and Act	28
12.C Atomic Action Descriptions	28
12.D Comparison of the Language Statistics	29
13 Benchmarks: Annotations and Baselines	32
13.A Ego-Exo Relation	32
13.B Ego-(Exo) Keystep Recognition	39
13.C Ego-(Exo) Proficiency Estimation . .	50
13.D Ego Pose	54



Figure 8. The Project Aria device used for egocentric recordings.

7. Aria Glasses

For the Ego-Exo4D project, we chose to use Project Aria devices [35]. Project Aria is an egocentric recording device in glasses form-factor created by Meta. It is designed as a *research tool* for egocentric machine perception and contextualized AI research, and available to researchers across the world through projectaria.com.

7.A. Device and Sensors

The Project Aria device is built to emulate future AR- or smart-glasses catering to machine perception and egocentric AI rather than human consumption. It is designed to be wearable for long periods of time without obstructing or impeding the wearer, allowing for natural motion even when performing highly dynamic activities—such as playing soccer or dancing. It has a total weight of 75g (compared to over 150g for a single GoPro camera), and fits just like a pair of glasses.

Further, the device integrates a rich sensor suite that is tightly calibrated and time-synchronized, capturing a broad range of modalities. For Ego-Exo4D, *recording profile 15* is used, which uses the following sensor configuration:

- **One rolling-shutter RGB camera** recording at 30 fps and 1408×1408 resolution. It is fitted with an F-Theta fisheye lens that covers a field of view of 110° .
- **Two global-shutter monochrome cameras** recording at 30 fps and 640×480 resolution. They provide peripheral vision, and are fitted with F-Theta fisheye lenses that cover a field of view of 150° .
- **Two monochrome eye-tracking cameras** recording at 10 fps and 320×240 resolution.
- **An array of seven microphones** recording spatial audio around the wearer.
- **Two IMUs** (800 Hz and 1000 Hz respectively), **a barometer** (50 fps) and **a magnetometer** (10 fps).
- **GNSS and WiFi** scanning were disabled for Ego-Exo4D for privacy reasons.

All sensor streams come with metadata such as timestamps and per-frame exposure times. All data is made available in raw form as part of the Ego-Exo4D dataset. For convenience, we also include pre-computed slices of data that suit specific purposes, e.g., 2D gaze points, mp4s of each cam-

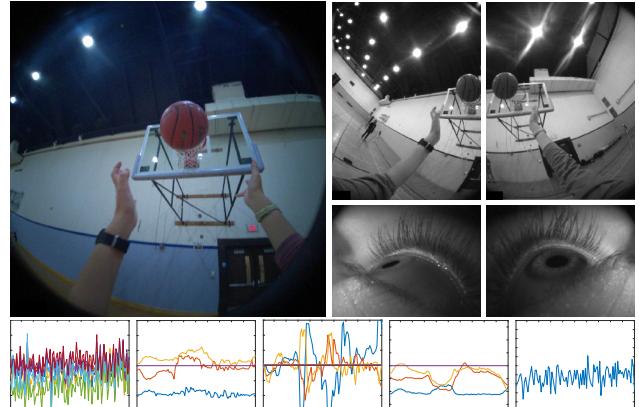


Figure 9. Sensor streams recorded by the Project Aria device. Top: RGB camera, left and right monochrome and eye cameras. Bottom: 10-second extracts from microphones, accelerometer, gyroscope, magnetometer and barometer respectively.

era, and smaller .vrs files with a subset of sensor streams.

7.B. Machine Perception Services (MPS)

Project Aria’s machine perception service (MPS) provides software building blocks that simplify leveraging the different modalities recorded. These functionalities are likely to be available as real-time, on-device capabilities in future AR- or smart-glasses. We use the following core functionalities currently offered by Project Aria, and include their raw output as part of the dataset. See [35] and the technical documentation for more details.

Calibration. All sensors are intrinsically and extrinsically calibrated. MPS also provides time-varying online-calibration that corrects for tiny deformations due to temperature changes or stress applied to the glasses frame.

Aria 6 DoF Localization. Every recording is localized precisely and robustly in a common, metric, gravity-aligned coordinate frame, using a state-of-the-art VIO and SLAM algorithm. This provides millimeter-accurate 6 DoF poses for every captured frame, as well as high-frequent (1 kHz) motion in-between camera frames.

Eye Gaze. The gaze direction of the user is estimated as a single outward-facing ray anchored in-between the wearer’s eyes. We use an optional eye gaze calibration procedure, where the mobile companion app directs the wearer to gaze at a pattern on the phone screen while performing specific head movements. This information was then used to generate a more accurate eye gaze direction, personalized to the particular wearer.

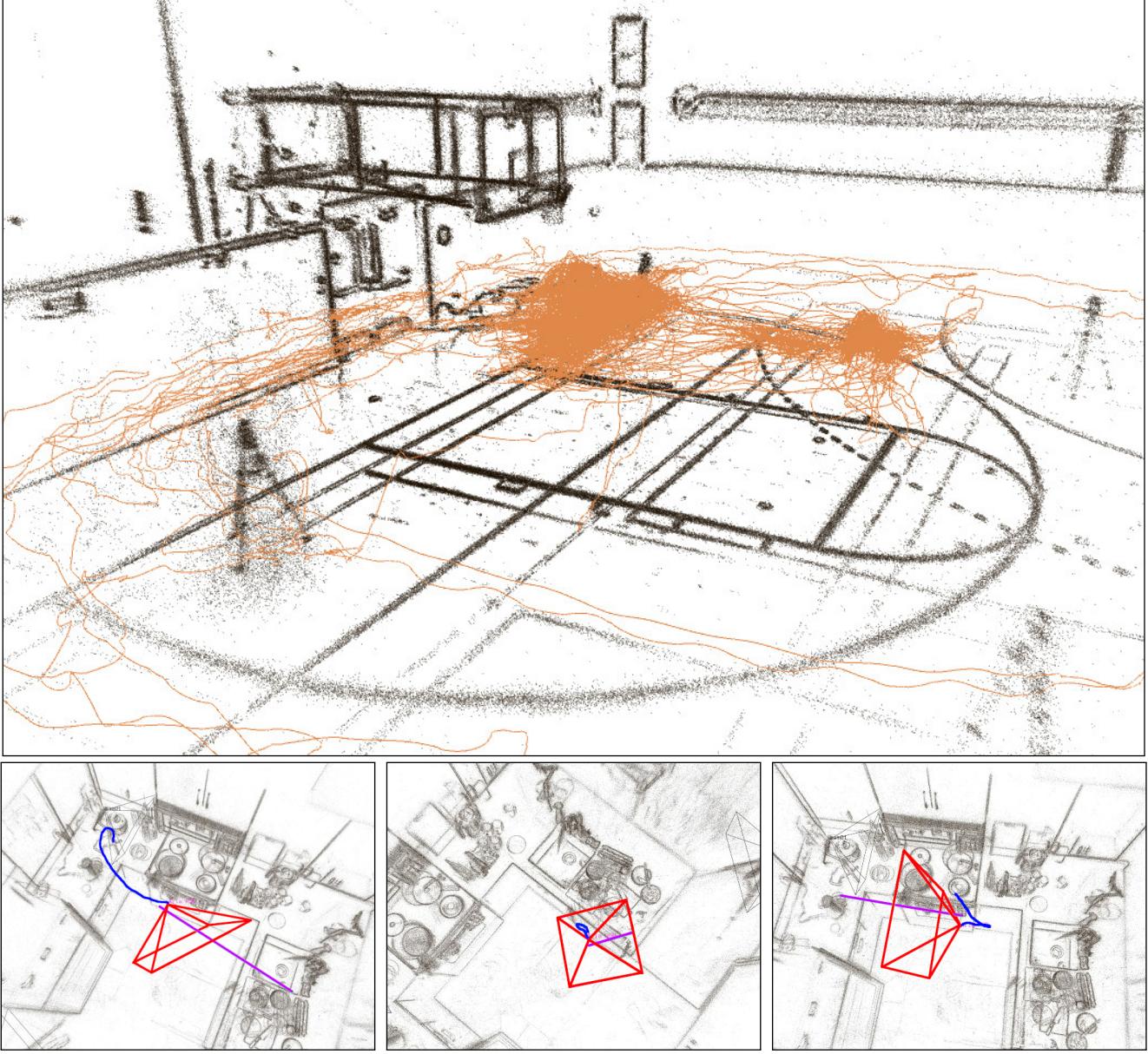


Figure 10. Aria MPS output for several recordings. Top: point cloud and estimated egocentric camera trajectory for a basketball session in Chapel Hill. This single continuous recording is 60 minutes long, has a total trajectory length of 2188 m, and contains 41 distinct takes. Bottom: three screenshots of a cooking recording, visualizing the current camera pose (red), eye gaze (purple), and last second of motion (blue).

Point Clouds. A 3D point cloud of static scene elements is triangulated from the moving Aria device, using photometric stereo over consecutive frames or left/right SLAM camera. The output contains both the 3D point clouds as well as the raw, causally computed, 2D observations of every point in the camera images.

GoPro 6 DoF Localization. For Ego-Exo4D, we added additional functionality on top of the existing Aria MPS

functionality, specifically to localize the static GoPro cameras. To achieve this, we use the map built with Aria’s SLAM cameras, and perform 6 DoF localization of GoPro frames on the map. To obtain the GoPro calibration, we manually calibrated one device in the lab to obtain default parameters, and then use the P4P [71] algorithm (with RANSAC to reject matching outliers) to estimate the 6 DoF pose, as well as re-estimate the focal length to compensate for possible calibration variation between devices.

7.C. Processing Summary

First, the MPS pipeline is invoked for *each full Aria recording*—these typically are about 20 minutes to 1 hour long and can include several takes, the hand-over in-between takes, as well as some other set-up steps. This is followed by localizing all GoPro videos of that scene as described above, and finally followed by time-synchronization across Aria and the GoPro cameras, as well as take-separation, as described below in Appendix 8.

There are total of 489 Aria recordings processed by MPS—containing the total 5,625 takes in the dataset. 95.9% of these recordings have successful Aria localization throughout the whole recording, with only 3.5% containing a partial tracking failure (leading to short gaps in the 6DoF trajectory). Three (0.6%) recordings failed completely. The most common failure reason is physical shock on the glasses, for example when the glasses are accidentally dropped on the ground or the table.

Furthermore we attempted to localize a total of 2,070 GoPro recordings, 91.4% of which are successfully localized. Similar to the Aria recordings, GoPro’s are localized on a *recording* level rather than on a *take* level. This helps in particular with very short takes as are common during physical activities—as there otherwise would not be sufficient visual overlap across Aria and GoPro perspectives. The most dominant reason for GoPro localization failure occurs when the GoPro is pointed to an texture-less area (e.g. a white table) which lacks the necessary visual features to perform localization. As the GoPro’s are static, this cannot be compensated for by device motion as is the case for the moving Aria device.

7.D. Tools and Ecosystem

Technical documentation and open-source tooling for Aria recordings and MPS output is available on Github² and the associated documentation page³. It includes both python and C++ tools to convert, load, and visualize data; as well as sample code for common machine perception and 3D computer vision tasks.

²https://github.com/facebookresearch/projectaria_tools

³https://facebookresearch.github.io/projectaria_tools/docs/intro

8. Camera Rig and Data Processing

The collection of ego-exo data at a global scale required us to develop a low-cost camera recording rig that was portable, auto-synchronized, and available internationally.

8.A. Hardware

Our unified camera rig is as follows: 1 Aria, 4 GoPros⁴, 1 GoPro Remote, 4 Tripods, 4 SD Cards, 4 Tripod Mount Adapters, 4 Velcro'd Battery Packs, 4 USB-A to USB-C Cables, 1 Glasses Sports Strap, 1 Smartphone, 1 Laptop or Tablet for questionnaires. The total cost excluding the Aria/phone/laptop is under \$3,000, with the majority of that going to the GoPros.

8.B. Time Sync

To sync cameras, we employ a pre-rendered sequence of QR Codes (*i.e.*, QR code video) that encode a wall-clock time. We show this QR code video using the smartphone at 29fps to all cameras in sequence and exploit the difference in frame rates to finely sync the cameras. In theory, the QR code decoded on a frame that captures a QR change is likely the one that was visible during that frame's center of exposure. With a single QR, the camera's center of exposure time could be anywhere within the 34.48ms that the QR is shown. However, with two consecutive frames with the same QRs, we can localize that time down to $\pm 0.574\text{ms}$. The same approach yields $\pm 0.558\text{ms}$ for the 59fps GoPros given 3 consecutive frames (see Figure 11), providing sub-frame synchronization accuracy.

An ‘audio sync’ fingerprint was played at the start and end of each capture to synchronize audio streams but has not been used.

Challenges and workarounds In practice, only $\sim 60\%$ of recorded captures yielded frame-accurate sync through our automated pipeline. Inaccurate sync causes included observed issues (*e.g.*, phone changing orientation mid-playback, video playback interruptions) and suspected ones (*e.g.*, videos not playing back at precisely 29fps, center exposure times not being evenly spaced). To recover these captures, we employed a manual sync procedure wherein people manually selected frame timestamps that should be aligned based on precisely time-localizable events, *e.g.*, a lighter first spark, a soccer ball making contact with a cleat, or a hand beginning a fast slide down the neck of a guitar. This unblocked the remaining $\sim 40\%$ of captures at the cost of less accurate sync.

⁴This represents the common core of the collection rig used in all capture settings. In certain captures, *additional* exo or ego GoPros are also used, as noted in Appendix 10.

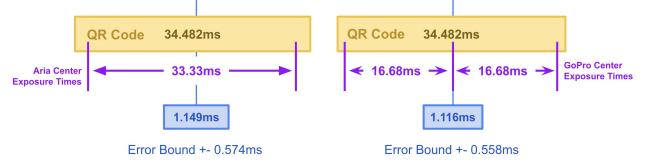


Figure 11. With a QR code timer playing back at exactly 29fps, cameras with evenly spaced center-exposures can be precisely time-localized to the QR timer with these multi-QR patterns.

Alternatives We explored and disqualified other sync options—notably using Timecode with TentacleSync or Ultrasync. Both of these solutions use LTC to encode a 1fps timestamp into the audio channel of a connected device. Using them with GoPros would cost us the stereo-audio modality, which we opted to keep to support audio-based research areas. We additionally lacked an ergonomic input solution for Aria to use while recording, so that mandated non-intrusive sync solutions.

8.C. Take Separation

To amortize the setup and tear down time required for each recording, we record multiple ‘takes’ (*i.e.*, one instance of a certain task) back-to-back and use a ‘Take Separator’ QR code (different from the time sync QR code video) that is identified in post-processing to auto-separate each take. This enables us to scale up recording—particularly for the physical scenarios where a single take can be less than a minute long. Data collectors track metadata for each take, identifying them by index and marking data such as participant ID (anonymous unique identifier), task (*e.g.*, making tea, making cucumber salad, performing CPR), and whether the take should be dropped (*i.e.*, if it is just setup time between activity enactments).

8.D. Recording Procedure

Our rig setup procedure entails setting up the stationary exo cameras in the recording environment and displaying QR codes to perform time sync and then take separations. Figure 12 overviews our recording procedure.

1. Position tripods, power on GoPros, and set camera angles to ensure maximum human coverage.
2. Begin Aria recording via smartphone.
3. Conduct a walk-around with the Aria glasses to build a basemap for 3D reconstruction and camera localization. Match the viewpoint of each GoPro camera by positioning the Aria directly in front of its lens.
4. Start QR Timesync Video off-screen. Show QR video to Aria RGB camera.
5. Use GoPro Remote to begin GoPro recording. Show QR video to each GoPro camera. Play Audio Sync fingerprint from the center of the space.

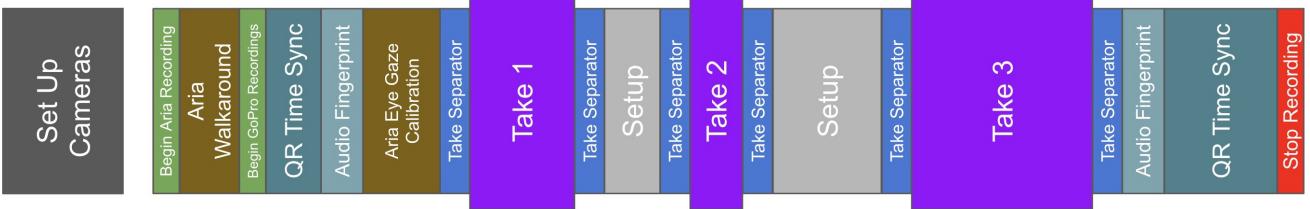


Figure 12. Overview of the recording procedure

6. Pass Aria glasses to (new) participant. Perform Eye Gaze Calibration via the Aria app. Show ‘Take Separator QR’ to one GoPro and begin the take. Show ‘Take Separator QR’ to one GoPro after the take is complete and repeat this step for each participant/take. Do not repeat gaze calibration if the participant has not changed.
7. Play Audio fingerprint from the center of the space. Restart the QR Timesync Video off-screen. Show it to the Aria RGB camera, then each GoPro. Stop recording on all cameras.

8.E. Extensions

The core camera rig was extended to handle onsite requirements and regional challenges. The team at Universidad de los Andes introduced a top-down (ceiling mounted) GoPro for dance, which was adopted by the team at the University of Pennsylvania with an overhead pole mount. The teams at University of Pennsylvania, IIIT-Hyderabad, and Indiana University added an additional egocentric, head-mounted GoPro.

Contribution Statement

Rawal Khirodkar proposed the hardware and software specifications for the ego-exo camera rig and helped design the capture protocol. Sean Crane investigated various hardware setups leading to the final rig configuration and helped draft the capture guidelines including recommended gear. Devansh Kukreja developed the sync and take separation algorithms, experimented with different equipment options (e.g., camera, timecode boxes, mount options), and designed the interface to transfer data; he also managed the ingestion pipeline, the collaboration with Aria, the integration of their code for EgoExo, and usage of the .vrs files.

Procedural	Physical
<p><i>Cooking:</i></p> <ul style="list-style-type: none"> - Omelette - Scrambled eggs - Tomato and egg - Sesame-ginger Asian salad - Greek salad - Dumplings - Noodles - Pasta - Sushi roll - Samosa - Coffee latte - Chai tea - Milk - Cookies - Brownies <p><i>Health:</i></p> <ul style="list-style-type: none"> - COVID test - Cardiopulmonary Resuscitation (CPR) <p><i>Bike repair:</i></p> <ul style="list-style-type: none"> - Remove/install a wheel - Replace an inner tube - Clean and lubricate the chain - Adjust rear derailleur (both limit screws & indexing) 	<p><i>Music:</i></p> <ul style="list-style-type: none"> - Violin - Piano - Guitar <p><i>Basketball:</i></p> <ul style="list-style-type: none"> - Mikan layup drill - Righthand reverse layup - Mid-range jump shot <p><i>Soccer:</i></p> <ul style="list-style-type: none"> - Freestyle dribbling - Freestyle juggling - Penalty kicks <p><i>Dance:</i></p> <ul style="list-style-type: none"> - Easy choreography - Advanced choreography

Table 1. Specific activities collected for the three Procedural and five Physical domains

9. Scenarios

Ego-Exo4D focuses on eight domains of human skill. See Table 1. These were selected to capture diversity, while also scoping our effort sufficiently to build a density of data within a core set of skills. Our scenarios were further targeted after a review of existing activity datasets, consideration of priority elements for transfer learning, and the identification of entry points to capture real-world scenarios available uniquely to our partners. To support data collection, we narrowed our scope to a small subset of tasks within each domain.

As discussed in the paper, we divide these into the categories of *procedural activities*—requiring a sequence of steps to achieve a goal state—and *physical activities*—requiring some degree of refined physical control to perform appropriately, but not necessarily following a step-by-step order. Note that in general an activity can be both procedural and physical; they are not mutually exclusive properties. Each consortium member selected multiple domains of focus from the eight targeted domains (cooking, health, etc.) based on their own preferences and local opportunities to capture data of these scenes at scale.

In working groups, our consortium developed data collection guidelines for each of our eight domains. These guidelines described the recommended camera positioning, instructions for participating camera wearers, along with important context-specific considerations. For example, given privacy concerns, our health guidelines required data collection participants to discard COVID tests before results were visible. These guidelines provided general parameters from which to collect data; however, they were not rigid steps.

To support diversity and implementation at a global scale, we allowed for site-specific adjustments to be made. For this reason, our dataset contains important nuances that should be noted, including, for example, differing standards for the implementation of CPR, cultural differences in the ingredients used for different targeted dishes, and location-specific bouldering routes.

Additionally, while data collection partners were encouraged to select scenarios of their preference, we did set broad targets for each site to collect at least three domains among the eight options. Since every domain is covered by more than one partner, the dataset exhibits visual variety from the different physical locations. Even within the captures done by a single partner, there are often multiple different sites used for filming (e.g., a couple different bike shops in the same city).

These domain selections were additional to our one cross-cutting scenario of cooking, which was collected by all partners. Cooking was a priority domain because it resonates around the world as a human need and interest. We are pleased to report that the cooking scenario of Ego-Exo4D contains more than 846 takes of cooking performed by more than 190 chefs in 64 different environments around the world, forming nearly 70 hours of ego video alone.

Figure 13 and 14 shows example frames illustrating the variety of the sites and tasks.

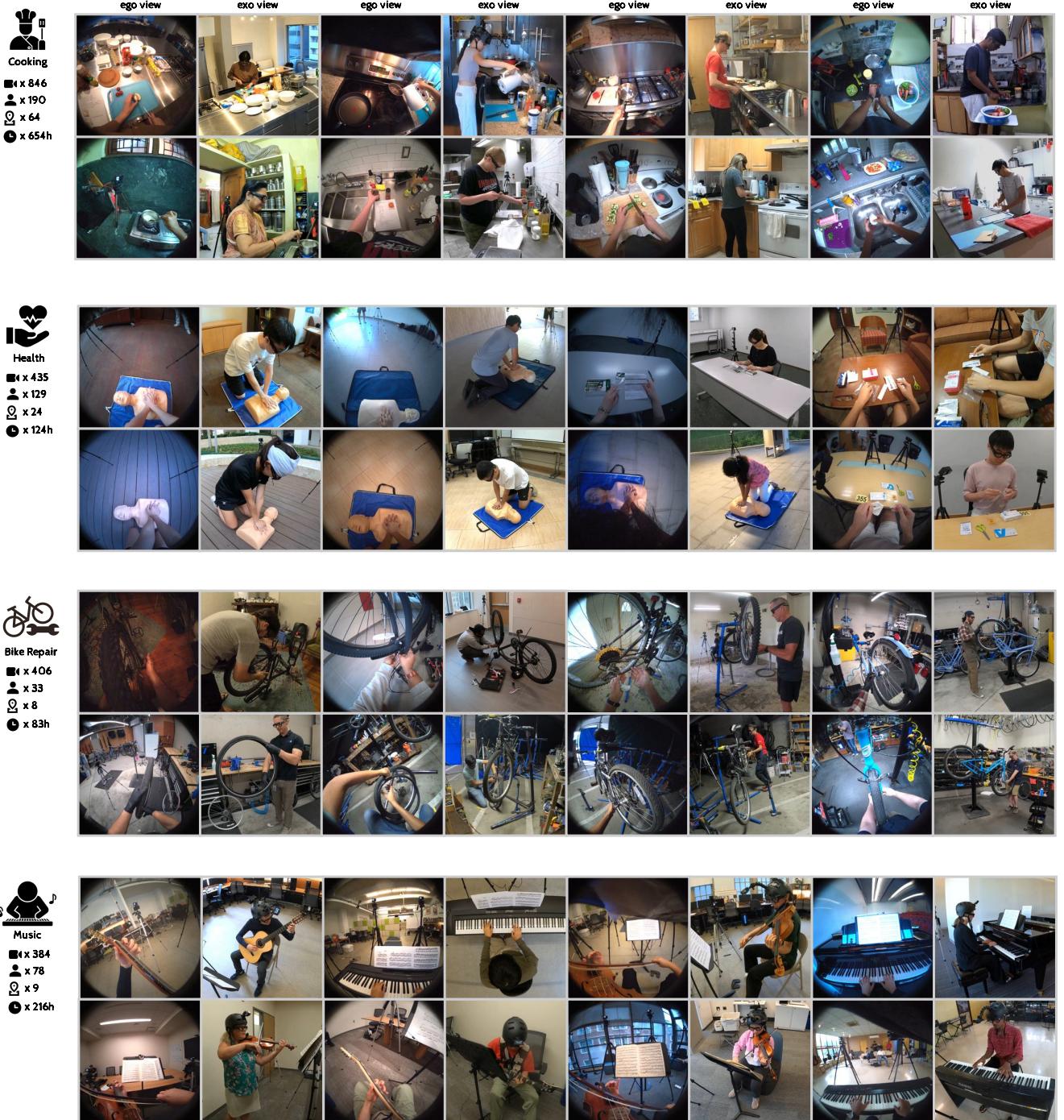


Figure 13. Ego-Exo4D captures skilled activity from 8 domains, in a wide variety of 131 scenes in 13 different cities in Japan, Colombia, Canada, India, Singapore, and seven US states. Every odd column shows an ego view, and the adjacent even column shows one of its paired exo views.



Figure 14. Ego-Exo4D captures skilled activity from 8 domains, in a wide variety of 131 scenes in 13 different cities in Japan, Colombia, Canada, India, Singapore, and seven US states. Every odd column shows an ego view, and the adjacent even column shows one of its paired exo views.

Contribution Statement **Bike Repair** Yale Song led the scenario development and collection guidelines for Bike Repair. Other contributors in the working group were David Crandall and Jim Rehg.

Culinary Manolis Savva led development of the scenario and collection guidelines. Other contributors in the working group: Andrew Westbury, Kumar Ashutosh, Deepti Ghadiyaram, Gene Byrne, Kristen Grauman, Santhosh Kumar Ramakrishnan, Shun Iwase, Yan Xu

Health Mike Zheng Shou was the lead author of the health scenarios guidelines.

Bouldering and Dance Pablo Arbeláez and Maria Es-cobar were the lead authors of the Bouldering and Dance guidelines.

Music Jianbo Shi and the team from the University of Pennsylvania were the lead authors of the Music guidelines.

Basketball Gedas Bertasius was the lead author of the Basketball guidelines.

Soccer Rawal Khirodkar was the lead author of the Soccer guidelines, camera placement, size of the capture area, and drills that were captured.

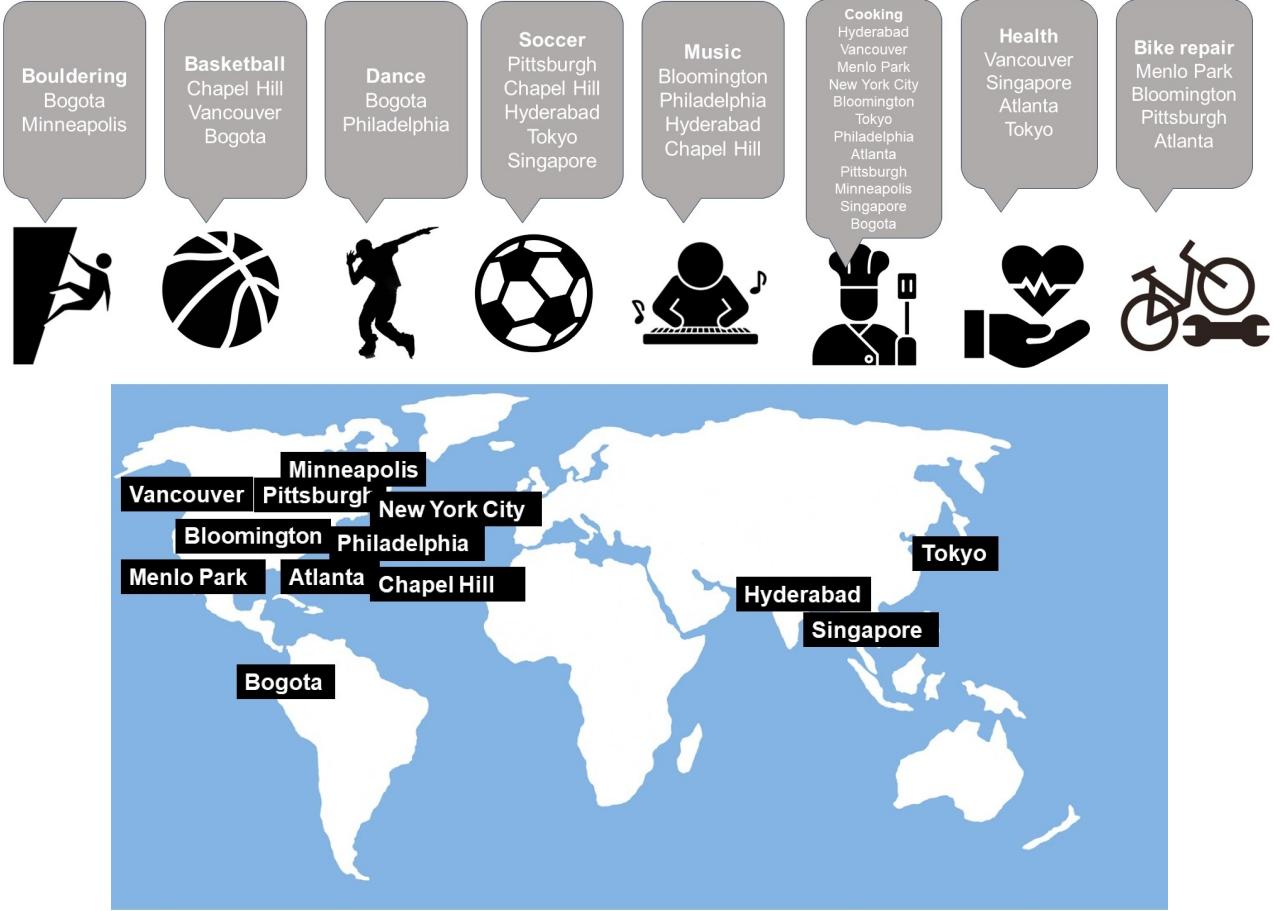


Figure 15. Geographic coverage of Ego-Exo4D and breakdown of which scenarios are captured in which cities. Note that even within a given city, there may be multiple sites (e.g., multiple bike repair shops or kitchens in the same city).

10. Data Collection

Twelve research labs came together for nearly two years to create Ego-Exo4D. Importantly, our collection across the sites was a coordinated effort, with common guidelines, scenarios, and camera rigs. In this way, the dataset is cohesive at the same time it is diverse. In this section we describe the data collection details that are specific to each partner site, e.g., how they recruited participants, which of the 8 scenarios they captured, or any modalities they added on top of the common rig.

Figure 15 shows the breakdown of which scenarios were captured by each partner institution as well as a map highlighting the locations of the 12 labs involved in data collection. Note that an additional four institutions not shown on the map are part of the consortium (e.g., contributing to benchmarks) but did not collect data. They are UT Austin (USA), KAUST (Saudi Arabia), University of Cata-nia (Italy), and University of Bristol (UK).

10.A. Carnegie Mellon University

Carnegie Mellon University focused on three skill-based activity scenarios: (1) soccer, (2) bike-repairs, (3) cooking. The exocentric cameras for our collections, four in total, were arranged approximately in a square configuration at a consistent height to capture the full range of the activity. Notably, for the soccer activities, an additional exocentric viewpoint was positioned inside the goal post to offer a more comprehensive perspective on the participants.

Soccer In the soccer scenario, we collaborated with professional players from the Pittsburgh Riverhounds team, representing the experts, and students from Carnegie Mellon University (CMU) as the beginners. We captured the soccer scenario across 4 different locations. The drills featured a variety of movements such as dribbling, goal kicks, and juggling, with each participant performing for a minimum of 3 minutes. This scenario resulted in roughly 4 hours of egocentric footage and 18 hours from exocentric perspectives, encompassing 32 participants in total.

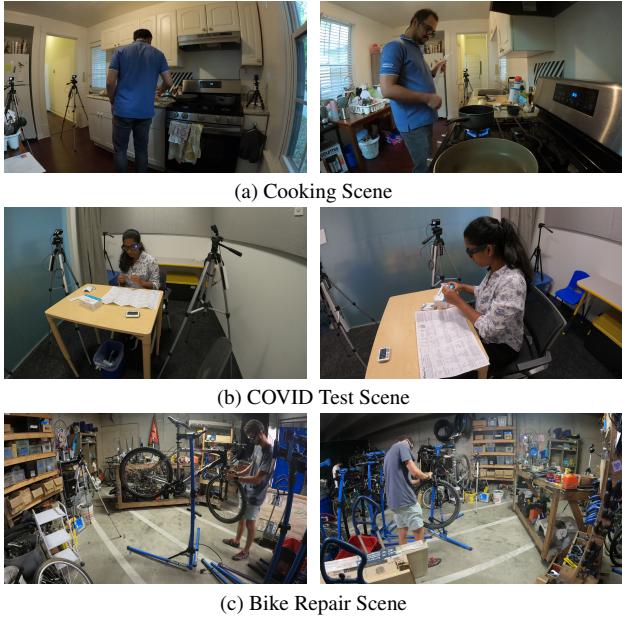


Figure 16. Views from two different cameras for each scenario collected in Atlanta, GA, USA.

Bike repair In the bike-repair segment, our experts were seasoned mechanics with over a decade of experience from Allegheny county. To ensure authenticity, we visited each mechanic in their respective shops to allow usage of their own tools and setup. Four tasks were captured for each bicycle, and we ensured bicycle diversity by selecting different sizes, shapes, colors, and makes. The tasks include tire removal, tube change/ inflation, tire reassembly, and clean/lube chain. This yielded 3 hours of egocentric recordings and 12 hours of exocentric footage, encompassing 22 different bicycles.

Cooking For the cooking section, we documented a professional chef in his traditional kitchen environment. Our dish of choice was scrambled eggs, and to inject variety, the chef prepared it using different techniques. This segment summed up to an hour of egocentric recordings and 4 hours from exocentric viewpoints.

All recordings were conducted in Pittsburgh, PA, USA, strictly adhering to CMU’s Institutional Review Board (IRB) guidelines. Every participant was briefed about the recording process, and prior to their involvement, a signed consent form was obtained.

10.B. FAIR, Meta

We collected 119 total takes of skills demonstrations in New York and three different locations in California. We focused on cooking and bike repair, taking advantage of the skilled workforce of chefs and bike technicians that serve our campuses and employees in these locations. We used the unified

camera rig of 1 Aria and 4 GoPros without any additional sensors.

Bike Repair Our skilled mechanics performed four different bike repairs for a total of 102 takes. We focused specifically on wheel repairs (removing and installing the wheel & flat repairs). While we strive for diversity in terms of the model of bikes, a majority of those in the dataset are drawn from standard fleet bike models, which contain identical parts and components. The location featured in the dataset is a well-equipped, industrial scale bike shop.

Cooking Our chefs recorded five different recipes as part of 17 unique takes. along with salads, egg dishes, and Asian Garlic Noodles. Locations featured in the dataset are three different professional kitchens used to prepare and serve hundreds of employees each day.

Internal documentation and processes ensured all participants provided informed consent to appear in the dataset and participation was strictly voluntary.

In total, we were able to mobilize five chefs and four bike mechanics. Due to workplace considerations, we did not collect demographic or age information about our chef and mechanic partners. Participating chefs and bike technicians are highly skilled with all research subjects reporting that they do the activity shown in the dataset daily or weekly. Similarly, eight research subjects have more than 10 years professional experience.

10.C. Georgia Tech

Collection at Georgia Tech focused on the Health, Cooking, and Bike Repair scenarios. Across these 3 scenarios, 279 takes were captured with 34 unique participants. For all scenarios, the unified camera rig was positioned such that 2 exocentric cameras would ensure capture of the participant’s hands, and the other 2 exocentric cameras would capture the participant’s full body and the full environment.

Participants were recruited from different sources including flyers, campus organizations, email lists, and word of mouth. Five of these participants completed data collections for 2 scenarios (4 participated in Health and Cooking, and 1 participated in Cooking and Bike Repair). Potential participants were provided with the study description and consent form prior to scheduling a recording session. At the beginning of each session, study personnel walked through the consent form with the participant, and answered any questions. The participant then reviewed and signed the consent form to confirm participation in the study.

The recording environment differed by scenario and included participants’ homes, campus meeting rooms, and an on-campus bike shop. Fig 16 shows a sample environment and camera setup for each of the Health, Cooking, and Bike Repair scenarios. Further details of the data collection specific to each scenario is provided below.

Health Participants for the Health Scenario took COVID rapid test kits while seated at a table. Recordings were captured in 2 different on-campus meeting rooms. Participants were recruited through campus email lists and flyers in local coffee shops. Each recording session lasted approximately 40-60 minutes and consisted of a participant completing 5-7 test kits, using 2-4 different types of test kits. 7 different types of COVID test kits were used across the full collection. In total, 96 takes were recorded from 16 unique participants.

Cooking Participants for the Cooking Scenario prepared dishes from three recipes: Asian Salad, Tomato & Eggs, and Garlic Noodles in their home kitchens. Participants were recruited via mailing lists of local apartment complexes, contacting participants from prior research studies, and word of mouth. Each recording session lasted 2-3 hours, capturing 3-6 takes of a recipe being cooked from start to finish. Participants cooked 2-3 of the recipes during their session, depending on dietary restrictions and preferences. Participants were provided with ingredients and a paper copy of the recipe, and used equipment from their own kitchen to prepare the food. In total, 71 takes from 15 unique participants were captured. The takes were about evenly distributed among the three recipes. Recordings were completed in 10 unique kitchen environments.

Bike Repair Participants for the Bike Repair Scenario performed repairs including taking off a wheel, putting on a wheel, replacing a tube, and cleaning a dirty chain. We recruited skilled participants from a campus bike repair organization. There were 8 unique participants, who each completed 1-3 recording sessions. Each session lasted 40-60 minutes and captured 5-7 takes of individual bike repairs. In total, 112 takes were recorded. showing the distribution across repair tasks. One session of 6 takes was recorded in a participant's home, while the rest were recorded in the campus organization's bike shop space, which is shown in Fig 16c. Due to the organization's access to a large quantity of used bicycles, there is large diversity in the make and model of bicycles across takes.

The study protocol was reviewed and approved by our Institutional Review Board (IRB).

10.D. IIIT-Hyderabad

In Hyderabad, we contributed to three scenarios - (a) cooking, (b) soccer, and (c) music. We formulated a data collection strategy tailored to the specific scenario, as outlined below.

Our primary objective was to comprehensively capture body and hand movements, along with their interactions with objects, during the execution of the activities. In general, we adhered to the standard camera setup instructions.

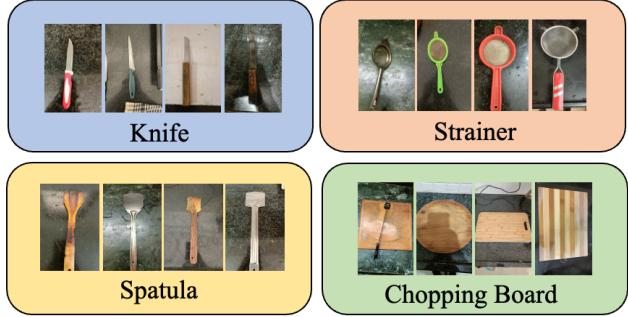


Figure 17. In Hyderabad, India, cooking was captured in different kitchens with socio-economic diversity. We observe that the same kitchen tools appeared in different shapes.

Nonetheless, we incorporated an extra exo-camera for capturing soccer activities in order to enhance the overall coverage of the event. Additionally, for music activities, we introduced a head-mounted Go-Pro camera.

This decision stemmed from the observation that expert musicians frequently do not directly look toward their instruments while playing. Consequently, the head-mounted camera guarantees continuous visibility of both the hands and the musical instruments, providing an ego-view perspective.

The collection in India was done during the peak summer, and this led us to a challenging situation where the cameras frequently shut down due to overheating. To address this, we mostly avoided capturing multiple takes in one capture and placed the cameras into an ice chest box in between the captures to cool them down.

Cooking For cooking, we reached out to people located in Hyderabad with varying socio-economic backgrounds and explained the data collection plan, and goals. We also requested them to engage their family members as well as friends in this data capturing process. Finally, we recorded the videos with the 41 informed participants capturing in a diverse set of 19 kitchens, geographically well-apart in and around Hyderabad, resulting in a rich dictionary of kitchen utensils (see Figure 17), narrations in four different languages. Additionally, we made an effort to ensure a balanced representation of genders in our overall data collection process.

Soccer For soccer, we reached out to three different soccer training schools in Hyderabad with the overall recording plan and process. They helped us in recruiting local soccer teams who play professional tournaments and practice almost everyday. We also recruited few players from our university soccer teams. In total, we recorded 49 participants, 'performing dribbling, juggling, and penalty-kick activities.

Music For the music scenario, we contacted one music school and recruited 4 musicians from them having at least 3 years of experience of playing either the piano, guitar,

or both instruments. To add diversity, the musicians were asked to play western as well as Indian pieces.

Our collection protocol was reviewed and approved by our university's Institutional Review Board (IRB). The primary conditions set forth by the IRB encompass the following aspects: (a) participants with 18+ age are deemed suitable for inclusion in the project, (b) participants have provided explicit consent for their facial and vocal presence to be featured in the released videos, (c) participants have willingly agreed to take part without receiving any immediate financial incentives from the videos, and (d) participants have the autonomy to engage in the activities in an environment of their choice. The participants were given the detailed descriptions of the project beforehand and requested to sign the consent form. Each participant received compensation as part of the process.

We selected participants from a wide range of age groups, spanning from 18 to 61 years old, to introduce an additional layer of diversity. Moreover, the participants were from diverse professional backgrounds (*e.g.*, coach, software engineer, data annotators, project managers etc.). Before sharing, we carefully examined each video to ensure there was no sensitive content.

10.E. Indiana University

We focused on cooking, bicycle maintenance, and music scenarios. All activities were collected using the unified camera rig, including additional sensors in specific scenarios. For cooking, the 4 GoPros were placed 90 degrees apart from each other, with 2 placed close to the participants to capture hands and objects and 2 placed further to capture the overall scene. In music, 4 GoPros were placed in front of the player, approximately 45 degrees apart from each other. In addition, we attached an additional GoPro HERO10 camera to the participant's head (using a helmet), tilted down roughly 80 degrees to capture hand movements. In bike repair environments, the 4 GoPros were placed 90 degrees apart from each other, of which 1 GoPro was placed close to the bike, 1 GoPro was placed close to the work-bench and tools, and 2 GoPros were placed further away from the participants to capture the overall scene.

Cooking For cooking, we had a total of 18 participants collect 72 takes and 20.5 hours of video. For 15 of the participants, we used a commercial test kitchen at our university. We purchased all of the ingredients and kitchen equipment ahead of time and had them ready when each participant arrived. We asked them to make four dishes (chai tea, sesame-ginger salad, tomato and eggs, and noodles) and provided printed recipes for these dishes. The remaining three participants chose to record in home kitchens, and the four dishes they made varied based on their preferences (one participant made omelet, cucumber salad, noodles, and chai tea, another made scrambled eggs, sesame-ginger

salad, sushi rolls, and brownies, and the third made scrambled eggs, cucumber salad, noodles, and milk tea). Due to concerns about food safety, we discarded (composted) the cooked dishes instead of allowing the participants to eat them.

Music For music, we had a total of 17 participants collect 60 takes and 6.5 hours of video. Participants were recruited based on their self-assessed proficiency in one of three instruments: piano, violin, or guitar. We recorded in 4 different locations including two studios, an office, and an auditorium that had a piano. Participants were instructed to play scales and arpeggios (2 mins), sheet music provided by us (3 mins), freeplaying (10 mins), and then recall and talk about any mistakes that were made during the playing and what could be improved (2 mins).

Bike repair For bike repair, a total of 13 participants recorded 108 takes and about 8 hours of video. We initially planned to hire professional bike technicians, but it was very difficult to recruit them in our relatively small city. Instead, we recruited more generally, looking for participants with (self-assessed) proficiency to do four basic bike maintenance tasks: removing a wheel, changing an inner tube, reinstalling a wheel, and cleaning and lubricating the chain. Most of the takes were recorded in a small house that is used for storage by our university's landscaping staff, and provided a realistic garage-like environment. We provided participants with a bike rack and supplies including bike tubes, pumps, tools, chain cleaner and lubricant, and gloves. To achieve diversity in different bikes and bike types, we asked participants to bring their own bike when possible, and we also provided 4 bikes (one of which belonged to one of the authors and the other three which we bought at a salvage shop). Most participants performed takes on about 3 bikes. One participant chose to record in an apartment, and one recorded in a hallway in a university building instead of the garage due to scheduling conflicts.

Our protocol was reviewed and approved by our university's Institutional Review Board. For each potential participant in each scenario, we first scheduled an online introduction meeting to tell them about the study and answer their questions and concerns. If they were interested, we agreed on the activity they would perform and when and where to meet for recording. We also sent them the informed consent form to give them sufficient time to review. On the recording day, we first asked them to sign the consent form, and then started recording their activities. All activities were recorded in an enclosed space to make sure that no one else accidentally entered the field of view of the cameras. We also ensured that the space did not have privacy-sensitive content, and we instructed participants not to use their phones or other devices that might show private content.

Within a few days, we securely sent the videos to the

participant so that they could review the video and ensure that they were comfortable sharing it with others. They also completed a brief online demographic study, and then were sent an incentive payment in the form of an electronic Amazon.com gift card. We made clear to participants that if they were not comfortable sharing their video, we would destroy it and they would still receive their incentive payment, although none of the participants chose this option. We gave the participants US\$20 in gift cards for each hour of their time spent recording (with a minimum of \$20, and partial hours rounded up to the nearest \$5). We gave an additional \$20 gift card to reimburse travel costs for those who came to our facilities to record (e.g. in our kitchen, bike repair shop, or on-campus studio or auditorium). For cooking and bike repair, we gave an additional \$20 gift card to participants who provided their own ingredients or bike maintenance supplies, to defray these expenses.

We recruited participants in the Bloomington, Indiana, USA area through online email advertisement, word of mouth, physical flyers, and posting on social media. We recruited participants who were 18 years of age or older, had self-assessed expertise in the activities as described above and could perform the tasks without wearing prescription glasses (which could interfere with the Aria's gaze tracking).

10.F. National University Singapore

In Singapore we focus on the following scenarios: soccer, health-related activities including COVID-19 ART testing and Cardiopulmonary Resuscitation (CPR), and cooking. In total, our collected data encompasses around 26 hours of egocentric videos and 117 hours of exocentric videos. These videos spread across 327 takes. In general, we adhered to the standard camera placement guidelines; however, for each scenario, we fine-tuned the position of the exo cameras based on practical considerations. For instance, in a small kitchen for cooking, we positioned the camera on the table to broaden its field of view.

Soccer For soccer, we conduct recordings at a university sports field. Our participants were primarily sourced through referrals provided by skilled participants recruited through online calls for participants. Additionally, during outdoor recording sessions, we occasionally invited surrounding bystanders to participate.

Health For health activities, we recorded in vacant classrooms, meeting rooms, and outdoor fields. CPR sessions are captured either in a yoga classroom or in a quiet, empty outdoor field. For recruitment, we circulated online calls for participants and then, for skilled activities like CPR, we collaborated with experts to organize training courses. Participants would participate in these courses and were trained to be proficient and then conducted recording afterwards.

Cooking As for cooking, which requires a kitchen, we

used the kitchen in our lab mates' apartments and arrange other participants to go there.

Our data capture has been approved by our university's Institutional Review Board (IRB). The main requirements include that participants: (1) agreed to take part in the study, (2) agreed to donate their speech, image, video, IMU, and 3D scan data for the purposes of this research, (3) agreed that their face, tattoos, and voice may appear in the data, (4) have the right to withdraw their recorded data at any time.

In Singapore, high temperatures often pose the challenge of camera overheating, particularly for GoPro cameras, which can lead to protective shutdowns and interrupt data collection. To mitigate this, we place small ice cubes wrapped in wet wipes on the GoPro cameras to help cool it down during recording. Furthermore, we attempted to schedule our participants' recordings in the evening or during an overcast day.

Our data pool comprises contributions from about 93 meticulously selected participants, ensuring a proficient completion of the recordings. Particularly in soccer, most participants have extensive experience and were members of their school or college soccer teams.

10.G. Simon Fraser University

We captured three types of scenarios in a variety of environments: cooking, basketball, and COVID-19 testing. In total, 88 participants carried out activities in the three scenarios we collected in a total of 61 data capture sessions, resulting in 519 activity takes.

We used the unified camera rig and followed the general collection guidelines with a number of small adjustments to facilitate scenario-specific capture. In kitchen and health scenarios where the participant interacts with small objects in tabletop height settings, the placement of exocentric cameras was optimized in a “two near, two far” setup to provide for visibility of the small objects and hands while also capturing the overall human pose during the activity.

Cooking The cooking scenario was captured in a decentralized fashion by going to the participants' own residences and asking them to cook in their kitchen. This allowed for diversity in the environment as well as in the participant during data capture. Our data capture sessions resulted in 112 cooking takes.

Basketball Collection for the basketball scenario was done in a “round robin” fashion to reduce player-to-player overhead. We targeted a spectrum of experience levels, for example going from university basketball team players who compete at the national level to more amateur basketball players who only have played basketball occasionally. We collected 355 takes of basketball activities.

Health Following the standard data collection guidelines for health activities, we gathered 52 takes of health activities.

We followed the institutional research board (IRB) process at our institution to acquire approval for the participant recruitment strategy, study setup, and participant consent acquisition forms. All participants consented to their data being collected and distributed for research purposes. Participants have the right to request that their data be withheld from inclusion in the dataset.

We recruited participants by word of mouth, reaching out to specific clubs and groups for some of the activities, and more generally through advertisement using university-affiliated communication channels.

10.H. Universidad de los Andes

We collected around 40 hours of video spanning four distinct scenarios that encompassed three physical activities (basketball, bouldering, and dancing) and one procedural activity (cooking). Figure 18 shows examples of the diverse scenarios that we collected. In total, we collected 2062 takes across all the activities. We used the unified camera rig with additional activity-specific sensors as described below.

Bouldering We partnered with a local climbing gym, which serves as a teaching and competition center in Colombia. We used the gym as the recording location and recruited participants who practice or teach bouldering there. Our focus was to recruit participants with four different levels of expertise: beginner, intermediate, advanced, and professional climbers. We hired expert route setters to design 33 climbing routes. These routes varied from beginner (V1) to expert level (V7). For data collection, each participant attempted to complete seven routes, having 3 minutes to make as many attempts as possible for each route. The routes were selected considering the expertise level of each participant. We located four exo cameras to capture each take; two horizontal cameras were facing the climbing wall, and the other two vertical cameras were on each side of the wall. Thus, the four cameras captured a complete view of the climbing wall and the participant's movements at every moment. We gathered 1251 takes for the bouldering scenario from 40 participants. We ensured ethnic, age, and expert-level diversity across the takes.

Dancing We collaborated with a salsa dance academy to use as a recording location and to help with participant recruitment. We recruited students from three expertise levels: beginners, intermediate, and advanced. According to the expertise level, each dancer performed different choreographies. Beginners recorded a single choreography, while intermediate and advanced participants recorded an additional one according to their expertise. Each attempt lasted one minute, and the dancer performed from six to ten attempts. The choreographies were designed by professional

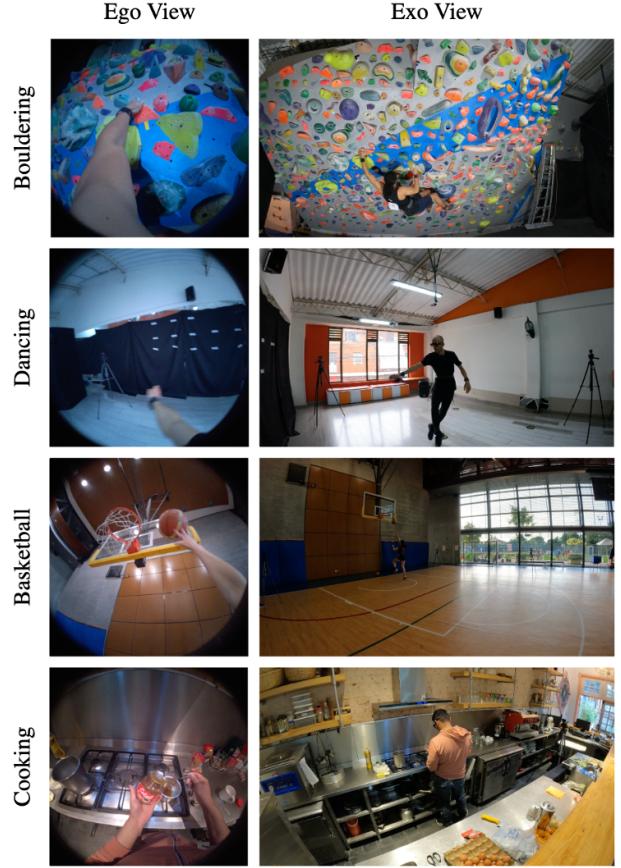


Figure 18. Egocentric and one exocentric view for each of the recorded scenarios in Bogota, Colombia.

dancers who teach at the academy. We used five exo cameras: four forming a square, defining the dancing area, and the fifth camera placed on the ceiling. Given the salsa dance's velocity and the movements' complexity, this fifth exo camera gave a crucial point of view for further analysis. We gathered 600 takes from 40 participants across the three expertise levels.

Basketball We collected data from the professional women's team and students from a basketball class at our University. Each participant performed six to ten attempts for each basketball exercise. We collected all captures at the basketball court at our University's Sports Center. For this setup, we used four exo cameras around the basketball ring, ensuring a complete view of each exercise. For this scenario, we collected 167 takes from 38 participants.

Cooking We rented a professional kitchen equipped with all the necessary utensils to perform the captures. We focused on collecting data from two types of recipes: a dish with egg and a drink. Each participant could choose be-

tween cooking an omelet, scrambled eggs, tomato and eggs, and coffee latte or tea for the drink. Each participant was free to choose how to complete each recipe. Thus, our takes show diverse ways to prepare each recipe. For this setup, we placed four exo cameras around the kitchen, all facing the user, to capture the whole kitchen without losing any detail of the person making the recipe. We placed two cameras on a counter facing the kitchen and the other two on each side of the kitchen. We collected 44 takes for the cooking scenario from 20 participants.

The Institutional Review Board (IRB) of our university reviewed and approved our study protocol. All participants signed a consent form before participating in the study.

We partnered with professional training centers for physical activities that helped us recruit volunteers with different expertise levels. These volunteers were previously familiarized with the activities and the environment where the captures occurred. In addition, we recruited family members, friends, and acquaintances of students and faculty members of our research group for cooking.

10.I. University of Minnesota

Collection at the University of Minnesota, Twin Cities focused on two main scenarios: Bouldering and Cooking. A total of 249 takes with 53 unique participants were collected. We collected all data using the unified camera rig with no additions.

Bouldering The bouldering activity was collected at a local bouldering gym, focusing on a wall with 14 different routes ranging in difficulty from beginner to expert. We collected 210 takes from 42 unique participants. Participants were asked to climb four to five routes of their choice, with the ability to take breaks within or between takes. Expert climbers who felt comfortable with the routes were able to narrate their approach and climb in real time. As participants were able to choose routes freely, our five exo-cameras were set up to accommodate the entire wall.

Cooking Cooking activity was collected on-site at each individual's home kitchen. Five exo cameras were set up in each kitchen to maximize coverage of both the participant and the environment. We captured 9 unique kitchen environments with 14 unique participants whose skill levels ranged from cooking novice to commercial chef. Participants focused on three recipes each (scrambled eggs, Greek salad, and pasta noodles from scratch), which were performed back-to-back on the day of recording.

Our data collection protocol was reviewed and approved by the Institutional Review Board at our university. At every take, the study personnel provides a guidance to a participant through the consent form prior to participation, ensuring the participant understands the purpose of the study and

all risks involved, with each participant receiving payment proportional to their contribution.

Participants were recruited via word of mouth, campus organizations, and digital flyers which were distributed via local social media (Facebook) communities.

10.J. University of North Carolina

Throughout our data collection at UNC, we focused on three skill-based activity scenarios: (1) basketball, (2) soccer, and (3) music drills. We used three unique environments (i.e., a basketball gym, a soccer field, and a music studio) to capture the data for each scenario. All recordings took place on the UNC campus. UNC's Institutional Review Board (IRB) reviewed and approved our study protocol. All participants signed a consent form before participating in the study.

To recruit participants, we used an online research study database, where participants from the local area could sign up to perform our study. We recruited participants willing to perform skill-based activities such as basketball, soccer, or music drills regardless of their skill level. Additionally, to recruit a more skilled group of participants, we contacted expert musicians from UNC's School of Music and athletes from UNC's basketball and soccer teams.

In total, we collected approximately 19 hours of egocentric and 76 hours of exocentric video data spanning approximately 548 takes of activity demonstrations from 56 participants (41 male, 15 female). Among the 56 participants, 44 were aged 18-25, 10 aged 25-50, and 2 aged 50-75. Furthermore, 26 participants had more than 10 years of experience in the scenario they chose to perform (e.g., basketball, soccer, music), 13 participants had 1-10 years of experience, and 17 had less than 1 year of experience. We used standard camera placement guidelines and the same recording devices described above.

Basketball All participants performed three basketball drills: Mikan Layup, Reverse Layup, and Mid-range Jump-shooting, for 388 takes. We recruited 11 expert players from the university team with 10+ years of experience. To improve the participant skill diversity in the dataset, we also recruited novice players with less than 1 year of playing experience. The location of data collection was a university basketball gym.

Music For the music scenario, we asked all 9 participants to play 5 minutes of scales and arpeggios and 10 minutes of free play for 27 takes. All of our participants were recruited from the university music club and considered themselves as experts at playing their respective instruments. The instruments featured in our collected dataset were piano, trombone, trumpet, and saxophone. The Ego-Exo4D music guidelines called for just piano, violin, and guitar, but we found it necessary to expand this list in order to gather data for this domain. All data was recorded in a

university music room.

Soccer For soccer, we focused on three drills: dribbling, juggling, and penalty shots for 133 takes across 12 unique participants. 7 of these participants were experts with 10+ years of experience, whereas the remaining 5 participants were casual soccer players. All videos were collected at a university soccer field.

Our study protocol was approved by the Institutional Review Board (IRB). All participants signed a consent form before participating in the study.

10.K. University of Pennsylvania

The University of Pennsylvania focused on capturing videos of experts of various levels playing musical instruments, dancing, and cooking. Over the spring and summer of 2023, UPenn captured 521 usable takes across 95 participants for the consortium’s collections with up to 7 views.

One primary goal of this project is to capture detailed body movement, especially hands, across ego view and exo views. We work to ensure highly engaged experts enjoy demonstrating their full skill capability.

The hand-object/instrument interaction region is the key to understanding human activities and evaluating their skills. Comprehensive hand pose information is especially important for the full analyses of scenarios collected at UPenn, especially the music scenario, where slight differences in finger motion result in entirely different performances.

We also observed that experts had a tendency to not need to look at their hands during play. Thus, we found the initial data capture using the general camera setup to lack crucial visual information in such scenarios due to:

- (1) (in ego view) limited field of view of Aria glasses, and skilled experts don’t need to look at their hands,
- (2) (in exo views) frequent occlusion and self-occlusion caused by participants’ motion.

We added two cameras to maximize the view coverage.

Head-mounted Camera: The head-mounted camera on a helmet angled downwards to capture the hand/body region: (1) (ego) it follows the subject’s body motion faithfully, and (2) (exo) it is designed to focus on the hand-object interaction region with much less self-occlusion. Empirically, we found this additional camera is crucial for capturing guitar, violin, and cooking scenarios.

Overhead Camera: We replace the head-mounted camera with an overhead camera in (1) piano scenarios, where the overhead camera can have similar performance, and (2) dance scenarios, where the helmet can dramatically worsen the experience and performance of the participants.

We believe the goal is not to maximize the number of hours captured but to have the participants show (1) diverse techniques to build models for the scenarios, and (2) unique techniques to demonstrate their skill levels.

To get the most representative recordings of the participants, we aim to maximize their engagement during the data capture. Specifically, we (1) walk through the whole process with the participants before the data capture to familiarize them with the setup (2) let them choose their favourite music piece to play or dance with in music and dance scenarios; and (3) have a narrate and act section for the musicians to demonstrate how they feel about their performance.

Music For musical instrument playing, classified as a “physical” activity, we captured takes of musicians (1) warming up (scales and or Etudes) (2) sight-reading simple sheet akin to Suzuki Practice books or Etudes exercises, and (3) freeplaying. We captured takes of violin, piano, and guitar, with a duet between a cello and a violin for one trial. Participants were recruited from a diverse pool of musicians, spanning the Penn Orchestra, local music schools, and independent music students. This pool’s experience ranged from professional instructors and performers to complete newbies. We totaled 275 takes over 37 participants. Notably, during the shooting, we observed that participants were particularly uncomfortable with the helmet used to mount the GoPro; it interfered with their head movements and the bow sometimes ended up knocking against the mounted GoPro. To combat this, we added additional cushioning to depending on the subject’s head shape and broke sessions into chunks to allow for breaks.

Cooking Cooking, categorized as ”procedural”, consisted of preparing four dishes: an egg dish, a salad, a noodle dish, and a dessert. The group of participants consisted primarily of Penn students with experience ranging from amateurs to hobbyists. Professionals were unavailable due to scheduling conflicts. We totaled 81 takes over 20 participants. The entire filming process was undertaken within a three-week span, primarily at the apartment of one of the team’s participants. This location expedited our data collection for this task by providing a stove and fridge for regular use.

Dancing Dance captures, classified as a ”physical”, consist of four takes of dancers performing dance routines to a song. The dance types recorded included Lindy-Hop Jazz, Bollywood, Latin, and Chinese Folk Dance; across these genres, we totaled 165 takes over 38 participants. The Lindy-Hop Jazz dancers came from the Jazz Swing Attacks, a dance club in Philadelphia. Contact was established via Instagram, and data, collected weekly over a month. This group contained a balanced mix of experienced instructors and beginner dancers. The Bollywood dancers, the Drexel’s Philly Maza, were recorded in the Drexel Engineering Building. They compete nationally but routinely train beginner recruits. The Chinese Folk Dancers were

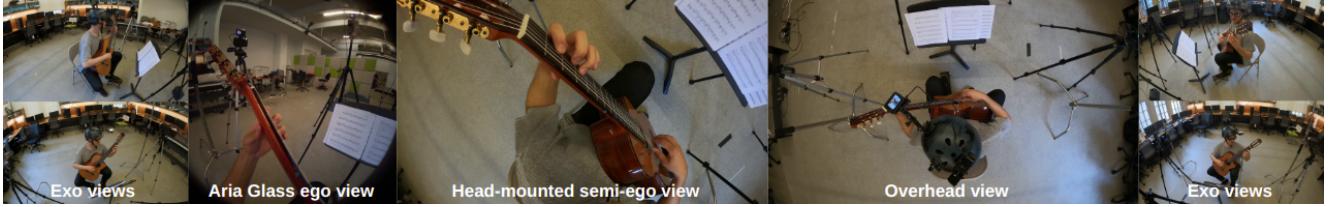


Figure 19. The Aria Glass ego view, head-mounted semi-ego view, overhead view and other static exo views in playing guitar in Philadelphia, PA, USA.

members of the local Great Wall Chinese School’s dance club and independent student volunteers with prior competitive dance training. These were captured in the SIG Lab for collections.

All participants were confirmed to be at least 18 years of age by the time of participation and gave written consent for participating in these data collection trials. The consent form, in compliance with IRB guidelines, but gives participants the choice to back out. The collected information on basic demographic information should not be used to identify participants individually. All other data collected per participant (prior experience with task, average times spent per session, etc) could not be used to identify participants.

10.L. University of Tokyo

In Tokyo, we collected video data for three scenarios: cooking and health for procedural activities and soccer for physical activities. We followed the standard camera configuration and calibration procedure of the Ego-Exo4D dataset for all scenarios. In the following paragraphs, we will describe the specifics of each scenario, particularly the unique aspects of our data gathering.

Cooking We recruited 12 Japanese participants living in the Tokyo area through a temporary employment agency. The gender and age of the participants were balanced to collect diverse behavior patterns. The participants cook three days or more each week in their daily lives. Each participant prepared three dishes: an omelet, a white radish & lettuce & tomato & cucumber salad, and a sushi roll. We recorded both versions with and without narrations for each dish and participant. A one-page summary of each recipe was provided before data collection and was shown during video recording so that the participants could prepare the dishes smoothly, and the procedure of each recipe should be consistent between the participants.

All video recording of the cooking scenario was done in a rental kitchen studio equipped with an island kitchen and all the necessary kitchenware for four consecutive days. The studio is situated in a busy location in downtown Tokyo, and some external noises, like ambulance sirens, were audible during the recordings. We collected 68 takes from 12 participants. Of the 68 takes, 34 takes are with nar-

tions, and 34 takes are without narrations. The length of each participant’s making each dish twice with and without narrations is about 35 minutes, ranging from 24 min 27 sec (Sushi roll) to 55 min 4 sec (Salad). The length includes time for the camera synchronization procedure.

During the recording of the cooking scenario, we discovered a flickering issue in some of the video data due to the incompatibility between the Aria Glass sampling rate and the power frequency in Tokyo. To overcome this issue, we attempted to shoot as much as possible in daylight and adjusted the fps when using artificial lighting. While processing the videos, we discovered some exo videos had decoding errors due to damaged frames. Each corrupted video contained one to three damaged frames for an unknown reason. To address this, we re-encoded these videos by replacing the damaged frames preventing decoding with the nearest good frame. Note that some videos still contain damaged frames as long as those frames did not influence decoding. In addition, the original MP4 files recorded by GoPro contain 4 streams: video, audio, data0, and data1, but the re-encoded videos only contain a video stream and an audio stream.

Health We recruited 17 Japanese participants living around Tokyo, Japan, through a temporary employment agency. The gender and age of the participants were balanced to collect diverse behavior patterns. We recorded videos of the 17 participants performing two tasks: COVID-19 rapid antigen testing using three test kits and performing CPR on a mannequin. We conducted all recordings in the same meeting room on campus over two days. For the COVID test, an instruction manual of each test kit was provided to the participants before and during the recording. Also, we did not show the participants or record any COVID test results for privacy protection. For CPR, the participants took an introductory lifesaving course provided by the Tokyo Fire Department before recording. Besides, a one-page summary of the CPR procedure was provided before the data collection and shown during the video recording. This is so that the participants can perform CPR smoothly and the procedure is consistent among the participants. We collected 73 takes from 17 participants. All of the CPR procedures (17 takes) were recorded without narration. For the

COVID test, we recorded the videos of the 17 participants using the three test kits (51 takes) without narrations. The video length of each participant’s performing CPR is 9 min 48 sec on average, including the camera synchronization procedure. Similarly, the video length of each participant’s using the three COVID test kits is 29 min 22 sec on average. Additionally, we recorded extra takes of 5 participants out of the 17 using a test kit with narrations (5 takes in total).

Soccer We gathered videos of 14 Japanese participants, each performing three fundamental soccer drills: dribbling and juggling for two minutes each and penalty kicks ten times. Of the 14 participants, 13 are soccer players from a university football club. We recruited them through the staff of the club. The remaining one participant is not from the club but is an expert with over ten years of soccer experience. All the participants are male, and their age ranges from 18 to 30s. We recorded the videos on an outdoor soccer field at a local university over four days, with three to four participants participating each day. For juggling, we instructed the participants to include various movements such as juggling with thigh, inside and outside of feet, and alternating feet. For penalty kicks, we instructed them to shoot to the right side of the goal 5 times and to the left side five times. During penalty kicks, a helper aids the participant in retrieving the ball. This helper stands within the goal area and might be recorded by some cameras. We collected 42 takes from 14 participants. All the takes were recorded without narrations.

Our university’s institutional review board reviewed and approved our study protocol. We explained the objective and the range of use of the videos through documents and took consent from each participant before the recording. In particular, we took the consent not to blur their faces to keep the naturalness of the videos.

Contributions Statement

Los Andes University Pablo Arbeláez - lead coordinator for data collection and collaborator on the overall project design; María Escobar - data collection for all phases, design of the collection setup and workflow, data inspection, ingestion, encoding, and metadata generation; Cristhian Forigua - data collection for all phases, participant recruitment, consent forms design, data inspection, communication with recording sites; Cristina González - data collection for phase 1, design of the collection setup and workflow, IRB management; Angela Castillo - data collection for phase 2, manual data inspection, and data analysis.

Georgia Tech James M. Rehg - lead coordinator for data collection and protocol design, and overall project manager; Bikram Boote - lead coordinator for data collection, including recruiting and ingestion; Fiona Ryan - contributed to data collection; Audrey Southerland - lead coordinator for

IRB development, contributed to recruiting.

National University Singapore Mike Zheng Shou - lead coordinator for data collection and protocol design, and overall project manager; Joya Chen - contributed to protocol design, camera setup design, data collection for all phases; Jia-Wei Liu - contributed to protocol design, camera setup design, data collection for all phases; Xinzhu Fu - contributed to data collection for all phases; Chenan Song - contributed to data collection for all phases.

Meta Andrew Westbury was the lead for data collection at our site, selecting scenarios, organizing capture sessions, recruiting participants, organizing and transferring data, and obtaining required approvals. In California, Hao Tang and Kevin Liang also supported all these functions, focused on bike repair. In New York, Devansh Kukreja and Alex Dinh lead collection for cooking scenarios. Miguel Martin also supported California-based collections and organized our local camera rig. Chefs Eton Chan and Dominic Ainza supported all culinary collections with technical guidance, recruitment, and coordination. Dimitri Elston coordinated and was the technical lead on bike collections. Adrian Salas supported pilot bouldering collections in California. Across all Ego-Exo4D collections, Devansh Kukreja continuously communicated and refined the recording procedure with universities, and problem-solved local recording issues.

University of North Carolina at Chapel Hill Gedas Bertasius - lead coordinator for data collection; Md Mohaiminul Islam - the main contributor to data collection and metadata processing across all scenarios; Wei Shan - contributed to data collection and metadata processing for the music and soccer scenarios; Jeff Zhuo - contributed to data collection and metadata processing for the soccer scenarios; Oluwatuminu Oguntola - contributed to participant recruiting and data collection for the music scenario.

Carnegie Mellon University Rawal Khirodkar developed the automatic 3D body keypoints extraction pipeline and collected a subset of the soccer, bike mechanic, and cooking sequences for the CMU portion of the dataset. Sean Crane was in charge of the data collection, IRB documents, capturing data, working with participants and processing the data for CMU. Abraham Gebreselasie ran the actionformer-based baseline for demonstration proficiency benchmark. Eugene Byrne served as the engineering lead for the initial design and implementation of the dataset, camera rig, processing pipeline and keystep annotations while at Meta. Subsequently at CMU, he assisted in the recognition benchmarks, implemented Ego-Exo transfer [79] (1 of the 3 baseline methods for keystep recognition) and the initial implementation of keystep action detection [184], and assisted in annotation/data quality generally.

Simon Fraser University Sanjay Haresh was the lead coordinator for data collection, including recruiting, data in-

gestion, and data analysis. Yongsen Mao also contributed to the data collection pipeline, recruiting, data ingestion, metadata annotations, and statistics computations. Manolis Savva advised on data collection, protocol design, and overall project management. We acknowledge the assistance of Hanxiao Jiang and Armin Kavian with recruitment and data collection.

University of Pennsylvania Edward Zhang led data collection efforts at UPenn and played a key role in subject recruitment, subject information collection, and on-site data recording. Jinxu Zhang led data management, information logging, and data transfer. Shan Su is the overall project lead, focusing on determining good camera configurations based on 3D reconstruction feasibility and fixing issues in data post-processing of time synchronization and take separation.

University of Tokyo Yoichi Sato served as the primary coordinator for data collection, while Ryosuke Furuta was responsible for data collection across all three scenarios, participant recruitment, and IRB submission. Zecheng Yu and Masatoshi Tateno provided support for data collection and were responsible for managing and transferring data. Takuma Yagi helped with the IRB submission process.

Indiana University David Crandall oversaw the overall effort at Indiana University, including protocol design and data collection. Weslie Khoo led the IRB protocol design and compliance, arranged logistics such as ordering equipment, and designed and oversaw the cooking scenario data collection. Yuchen Wang and Ziwei Zhao co-led the participant recruitment and data collection for all three scenarios. Ziwei Zhao led data preparation and transfer. We also acknowledge Manasi Swaminathan who assisted with data collection and video synchronization.

IIT-Hyderabad Avijit Dasgupta was the lead on the ground for data collection in Hyderabad helping in organizing capture sessions, data collection, and managing and transferring data. Siddhant Bansal helped in the early stages with IRB application, consent forms, and pilot studies. C. V. Jawahar was the lead coordinator for data collection helped in selecting the scenarios, and recruiting the participants.

University of Minnesota Hyun Soo Park oversaw the overall effort at the University of Minnesota, Twin Cities, including protocol design and data collection. Zachary Chavis led the IRB protocol design and compliance, arranged logistics such as ordering equipment, participant recruitment, and designed and oversaw all scenarios of data collection. Anush Kumar assisted data collection for all scenarios.

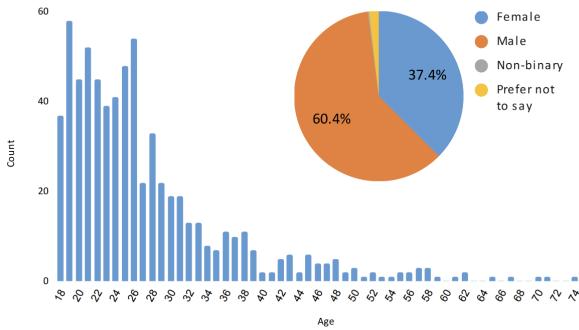


Figure 20. Participants' self-reported demographic information

11. Participants

Next we overview the background of the participants (camera wearers) in the Ego-Exo4D dataset.

11.A. Demographics

We provide self-declared information on ethnic groups by the participants. Sharing this information was optional for all research subjects. Ethnicity is reported based on location specific categories as defined by the relevant partner lab. No such information was gathered from research subjects participating in our collections in California, New York, and Pittsburgh, Pennsylvania. Aggregated gender and age information of all participants in Ego-Exo4D is provided in Figure 20.

Atlanta, Georgia, USA 100% of participants that reside in Fulton County, Georgia self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
Asian	23
White	8
Hispanic/Latino	3

Bloomington, Indiana, USA 100% of participants that reside in Monroe County, Indiana self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
Asian	22
Black	1
Middle Eastern	1
White	18
Prefer not to say	1

Minneapolis, Minnesota, USA 100% of participants that reside in Hennepin County, Minnesota self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
White	41
Hispanic/Latino	4
Asian	8
Black	1

Tokyo, Japan 100% of participants that reside in Tokyo self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
Asian (Japanese)	45

Hyderabad, India 100% of participants that reside in Hyderabad self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
Asian (Indian)	95

Chapel Hill, North Carolina, USA 100% of participants that reside in Orange County, North Carolina self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
White	20
Indian	1
Asian	13
African American	9
Hispanic/Latino	3
Prefer not to say	3

Vancouver, British Columbia, Canada 100% of participants that reside in Vancouver self-reported their ethnic group membership as follows. Please note that research subjects in this case opted not to use any assigned category and independently described their identity.

Ethnicity	Number of participants
African/Nigerian	4
Asian	9
White/Caucasian	10
Chinese	26
European	1
Iranian/Persian	14
Italian	1
Jamaican	2
Kazakh	1
Kyrgyz	2
Middle Eastern	1
Mixed	3
South Asian	2

Philadelphia, Pennsylvania, USA 100% of participants that reside in Philadelphia Country, Pennsylvania self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
White/Caucasian	10
Asian	30
African American	3
Hispanic/Latino	4
Prefer not to say	43

Singapore 100% of participants that reside in Singapore self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
Chinese	65
Indian	3
Singaporean	2
Indian/Chinese	2
Prefer not to say	17

Bogota, Colombia 100% of participants that reside in Singapore self-reported their ethnic group membership as follows:

Ethnicity	Number of participants
Black/ Afro-descendant/ Afro-Colombian	7
Mixed	104
Palenquero	1
Raizal	1
White/Caucasian	38
Prefer not to say	23

11.B. Participant Surveys

To ensure consistent, high quality annotations (which will be discussed in Appendix 13), we identified three aspects from data collection where additional information from the participants’ recording is crucial. These are: the *skill* of the participant, the *objects* they are using, and the *actions* they are completing. The three aspects represent information that either cannot be captured by other annotators after the fact or could result in noise within the annotation process. Accordingly, we designed participant surveys to cover key information regarding the participants’ *skill*. To capture the *actions* and *objects* with which they interact, we ask participants to perform a round of first-person narrations called “narrate-and-act” (cf. Appendix 12.B).

Participant surveys were separated into two: a pre-task questionnaire and a post-task questionnaire. The pre-task questionnaire aims to capture the participant’s perceived skill level whereas the post-task questionnaire captures the

participant’s reflection on how well the task went. The list of questions for both questionnaires can be found in Table 2 with questions/answers designed for consistency and ease of filling in, as participants would be filling these out before/after each recording. This involved using multiple choice and Yes/No answers with open text fields being utilized sparingly.

Pre-task Questionnaire Within this questionnaire, the aim is to capture the participant’s perceived skill level regarding the task that they are about to perform. To maximize consistency across participants, we forgo asking participants to self-rate themselves as experts, novice, etc.—instead, we ask participants more quantifiable questions such as length of time the participant has been doing this task and frequency.⁵ Additional questions regarding whether the participant has taught the task to others and whether they have recorded/watched videos about the task also give signals regarding the proficiency of the user. Note that the information gathered is scenario- and take-specific, since the number of hours to be considered an expert or amateur can be wildly different (e.g. cooking vs. bouldering).

Post-task Questionnaire The post-task questionnaire focuses on capturing details of the take and how it went according to the participant. Questions are asked about mistakes, objects, and time taken compared to predicted time, which is beneficial for downstream annotation.

The survey data will be shared with the dataset.

⁵We also obtain proficiency ratings for the participants via our expert commentators (cf. Appendix 12.A).

	Question	Answer Type
Pre-Task	Recording Location	multiple choice
	How many times do you estimate you have done this task?	multiple choice
	How often do you carry out this task?	multiple choice
	How many years have you been doing this task?	multiple choice
	Have you taught this activity to others before?	Yes/No
	Have you recorded a video of yourself carrying out or explaining this task before?	Yes/No
	Have you watched videos of others doing this task before?	Yes/No
	Do you have any qualifications/professional training that are related to the task?	Yes/No
	How long does it typically take you to complete this task?*	text
	How long would you typically spend in one practice session of this task?†	text
Post-Task	Self Reported Quality	multiple choice
	Completed Route?‡	Yes/No
	What mistakes/errors did you make during this task?	text
	Any issues with the familiarity of the tools/location?	text
	Did it take longer/shorter than your initial expectation and why?	text
	How did you find wearing the camera?	multiple choice
	How easy was the setup for recording?	multiple choice
	Any other comments to take on board?	text

Table 2. Questions for the pre-task and post-task questionnaires. *: Only applicable for non-dance/non-music scenarios. †: Only applicable for Dance/Music scenarios. ‡: Bouldering scenario only.

Bike Repair



Narrate and Act: It's nice and tight, clockwise to tighten that drive side nut, do another quick check.

Atomic Action Description: C tightens the axle nut with the bicycle skewer in his right hand.

Expert Commentary: Now the mechanic is applying pressure to the left side of the wheel and pushing the wheel towards the right. That allows to have the cog pull the chain and create a little more tension on the chain. Meanwhile, he is tightening the right side nut and making sure that there is some tension on the chain allowing a little slack but also holding it in place.



Narrate and Act: Bike Repair

Atomic Action Description: Bike Repair

Expert Commentary: Bike Repair

Narrate and Act: It's nice and tight, clockwise to tighten that drive side nut, do another quick check.

Atomic Action Description: C tightens the axle nut with the bicycle skewer in his right hand.

Expert Commentary: Now the mechanic is applying pressure to the left side of the wheel and pushing the wheel towards the right. That allows to have the cog pull the chain and create a little more tension on the chain. Meanwhile, he is tightening the right side nut and making sure that there is some tension on the chain allowing a little slack but also holding it in place.

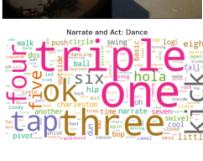
Dance



Narrate and Act: And you want to move very slowly in between these poses.

Atomic Action Description: C steps sideways while moving her hands in a circular motion.

Expert Commentary: For the last few counts, the dancer's arm movement has been very good. It's been very fluid, alternating one hand higher and one hand lower.



Narrate and Act: Dance

Atomic Action Description: Dance

Expert Commentary: Dance

Narrate and Act: And you want to move very slowly in between these poses.

Atomic Action Description: C steps sideways while moving her hands in a circular motion.

Expert Commentary: For the last few counts, the dancer's arm movement has been very good. It's been very fluid, alternating one hand higher and one hand lower.

Health



Narrate and Act: Squeeze the swab in the tube.

Atomic Action Description: C adjusts the testing plate on the table with his left hand.

Expert Commentary: He does get the swab into the solution tube fairly quickly after finishing collecting the specimen and that is what you want to do with this. You'd like to get the specimen as quickly as possible from the nostril into the tube. Any testing material that was available in there is going to be best tested as quickly as possible.



Narrate and Act: Health

Atomic Action Description: Health

Expert Commentary: Health

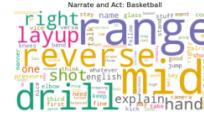
Basketball



Narrate and Act: Alright, so for a mid-range shot, you want to stay nice and low, keep your elbow tucked.

Atomic Action Description: C shoots a jump shot at the hoop with both hands.

Expert Commentary: Awesome high release here, way above his head. This is great. Elbows nice and tucked in. He's got a little bit of bend on the knees.



Narrate and Act: Basketball

Atomic Action Description: Basketball

Expert Commentary: Basketball

Figure 21

Figure 21. Examples of the three different annotation styles: narrate and act, atomic action descriptions, and expert commentaries from four of the scenarios (bike repair, health, dance, and basketball). We also include word clouds which highlight the differences in vocabularies per scenario. In narrate and act text, we see how the participants briefly describes what they are doing and why, whereas the atomic action descriptions provide strictly a statement about the visible actions. The expert commentary offers an expert's critique of what is shown, commenting on strengths and weaknesses and explaining how the participant's actions affect their performance.

12. Language Descriptions

As introduced in the main paper, Ego-Exo4D provides three forms of parallel text corpora for the video: expert commentary, narrate and act, and atomic action descriptions. Figure 21 and Table 3 show examples from different scenarios highlighting their distinctions in style and point of view. In the following we elaborate on their protocols, show examples, and analyze their differences.

Ego-Exo4D's language annotations are not targeted at any particular benchmark task or taxonomy. They aim to be a general resource that will inspire new language-vision possibilities, such as learning how to generate coaching or teaching advice given a demonstration video, learning how to spot skill and lack of skill, or learning general multimodal embeddings in the procedural and skilled activity domain. We also anticipate the temporally grounded descriptions to be valuable for pre-training foundation models [82, 119] or automated video captioning [56, 107, 194]. Furthermore, the time-anchored aspect of the three language annotations provides the opportunity to retrieve time points in specific Ego-Exo4D videos that correspond to queried moments, actions, or phrases. Finally, they are valuable to mine the dataset for the distribution of objects and activities present, e.g., for taxonomy formation.

12.A. Expert Commentary

Ego-Exo4D offers a new video-language resource specifically designed to support the study of human skill. We mo-

bilize 52 expert commentators, each with rich qualifications and skills in one of our focus domains, to provide time-anchored critiques and explanations of all the skill demonstrations in the dataset. Specifically, our expert commentators use spoken word to describe what is most effective or ineffective about the camera wearer's actions, review the quality of the execution, and identify mistakes. Experts also provide an overall proficiency rating on each skills demonstration, assessing how well the task was performed with a short written justification. The following section describes the qualifications and background of our experts; commentary instructions and tooling; and statistics on collected commentaries to date.

Experts and Expert Qualifications Over a period of three months, we recruited experts from across each of Ego-Exo4D's eight core domains, working with a selection panel of ten individuals to review qualifications and engage in more than 68 interviews. Our selection criteria focused on technical skills, communication, and task completion along with performance reviews during a live video commentating exercise. Among the numerous candidates we interviewed, we were pleased to select 52 to join the effort.

On average, 90% of recruited experts possess more than 10 years of professional experience and all have served during this time in the capacity of a coach, instructor or mentor. All experts further have either an advanced degree in their domain of focus or an industry certification. Certi-



Proficiency score: 10

Commentary: Great footwork. He's using dribble to set up his footwork and his shot. Stepping onto that left foot bringing the ball. I love that his eyes and head are up. He already knows where he's going to go.



Proficiency score: 5

Commentary: The dancer's hands should be a bit higher. This line should be completely straight in front of him at his shoulder length. It shouldn't be beginning to dip lower.

Figure 22. Two examples of expert commentary and proficiency scores, along with spatial drawings (red) done by the expert to augment the spoken comments.

fication authorities include the US Soccer Federation, the American Culinary Federation, USA Climbing, the American Red Cross, Trek Bikes, and New York State’s Initial Certification in Teaching Dance, among others.

Multiple individuals were recruited across each domain, with the goal of generating language and expertise diversity. Table 4 shows the number of expert contributors per domain. Due to employment considerations, all experts are residents of the United States.

Commentary Instructions and Tooling Experts are provided with two time-synchronized videos of each Ego-Exo4D skills demonstration—one showing the egocentric view and another providing a single exocentric perspective specifically selected by annotators as the view that provides the best visibility on the scene. Experts are then asked to watch the video in full without commenting to gain an understanding of the skills demonstration and plan out important points to note in their commentary.

After this initial viewing, experts are asked to provide a numeric proficiency score on the scale of 1 (least skilled) to 10 (most skilled), rating how well the camera wearer performed the task. A short written description of why the chosen score was selected is also provided at this stage. In many cases, experts coordinated within their domain group

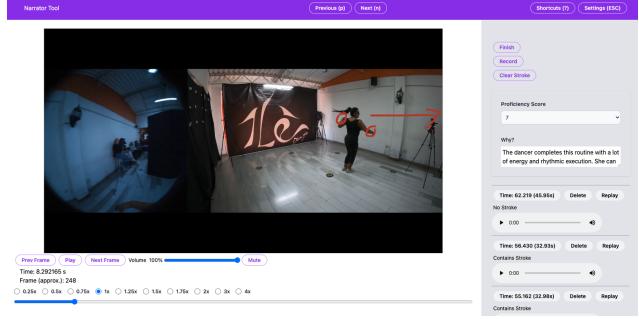


Figure 23. Our expert commentary web tool called *Narrator* provides an easy-to-use platform for experts. Experts can stream video, record audio commentaries, and provide proficiency ratings and justifications. The tool also supports drawing on the video feed (see red arrow and circles on the right frame), allowing for manual spatial grounding during commentary.

to calibrate this scoring. While each proficiency rating is at the discretion of individual experts, domain-level calibration provided a general framework for assessing skills demonstrations.

With proficiency scores noted, experts play the video from the start and pause at critical moments of the demonstration to record verbal descriptions of what is observed to be most effective or ineffective about the camera wearer’s actions, offer performance tips, identify mistakes or suboptimal execution, and why they matter. In all cases, commentary is time-anchored and retrospective, focusing on insights and perspectives relating to actions visible to that point in the video. We choose to collect commentary as verbal recordings in order to maintain the naturalness of the performance descriptions and do so quickly. In all cases, experts had the option to use a “telestrator” tool to enhance their commentary with freehand sketches to spatially localize information or otherwise help explain a point. See Figure 22. We transcribe the commentaries automatically with OpenAI’s Whisper for automatic speech recognition [121]. Table 3 (right column) and Figure 21 shows example commentaries. In total, our experts have dedicated more than 6,000 hours of effort to date to provide this commentary.

To collect expert commentary, we developed a web-based tool, which is open sourced as part of the Ego-Exo4D dataset and benchmark suite. Known as the Narrator, this application supports video playback for Ego-Exo4D skills demonstrations, records time-stamped verbal commentary, and allows exporting and viewing commented videos. As a web-based platform, the Narrator can be simply accessed through a browser, with minimal set-up and less restrictive system requirements compared to tools requiring local installation. These attributes made it efficient to onboard and manage our geographically distributed experts. We acknowledge the EPIC Narrator [25] as the open-sourced in-

Domain	Atomic Action Description	Narrate and Act	Example commentary
Cooking	C turns on heat on the gas burner.	So I'm going to start out by boiling some water.	Here the preparer is checking the pasta for done-ness. It's important to do this and not rely on what a package says. Use a package that gives you cooking time as a guideline and start to check your pasta, you know, a few minutes before the maximum amount of time given for cooking that specific pasta.
Health	C inserts the nasal swab in the buffer test tube on the covid test kit pack with his right hand.	Open the newly picked up tube, place the swab in the tube, stirring the swab in the tube.	So this individual has done a great job of making sure that her nasal passageways have adequate time in contact with her nasal swab. Something that might make it a little bit easier for her is if she could tilt her head back just a bit so that she wouldn't have to strain quite so much to get that access. Additionally, she did a great job making sure that the nasal swab was about an inch into her nose.
Bike repair	C holds the bike wheel with her left hand.	And then I will locate the location of the valve cap and pull the tube out of the wheel.	It's a great method to always double check or do a pre-check before beginning work on a bicycle to make sure the issue that you are working to fix is the only issue that is occurring. If not, you could find a secondary issue or something else that may be greater than the one you are currently working on.
Music	C puts the bow on the violin with his right hand.	So regardless of how tricky left hand passage work is you want to always keep your bow completely independent.	This is a really great use of the bow and decision to play in this middle third of the bow. This is exactly where they should be playing. And we can hear that the note envelope is very consistent and that it's very controlled and that it also allows the rhythm to be stable...
Basketball	C runs towards the hoop with the basketball.	Now I'm going to do a reverse layup, stepping right, left, going up with the right hand.	As the ball goes through the basket, she catches the ball and does an excellent job of keeping the ball high, never allowing the ball to drop down to her waist area, but keeping the ball high in her upper chest, neck area throughout the drill...
Bouldering	C places both hands on a red hand hold.	So I know that a lot of these holds, I'm going to need my weight leaning to the left to utilize	Once the climber recovered from the foot cut, the climber pasted the right foot on this foot jib and then did a toe match. So brought this foot in and then dropped the right foot down and to the right to again counterbalance so that the climber can then move their left hand out left. But at this point the climber is just a little too gassed to be able to make this move, which is unfortunate.
Soccer	C kicks the ball to the right with his right foot.		Angle approach, start position is good, maybe slightly squarer than 45, but again because the intended outcome from previous actions is into the left, by being a little squarer is going to help him be able to rotate his hips to move to the left, but on a slight angle is good and help him with his technical action.
Dance	C moves her right leg forward while swinging both hands.	Ring and wing, one, two, one, two, three	She is doing these steps in place when she's traveling forward. At this point, she really could be further forward all the way, still on the screen, but towards the edge of the screen, if she was to take bigger steps. And she could take bigger steps if she bends her knees and lowers her center of gravity and then extends her leg outward...

27
 Table 3. Example excerpts from all three language types. Experts are charged with critiquing the performance of the participants, pointing out strengths and weaknesses and explaining how the participant's approach influences the quality of their skill demonstration. Narrate and act focuses more on what the camera wearer is doing and, sometimes, briefly why. Atomic action descriptions are about the specific actions seen.

Domain	# experts
Basketball	6
Music	6
Soccer	5
Dance	6
Bouldering	6
Health	5
Cooking	11
Bike Repair	7

Table 4. Number of experts providing commentary in each domain

spiration and source code for this initiative.

Commentary Analysis We apply the expert commentary to all video clips in the dataset, and we ask at least 2 and as many as 5 distinct experts to commentate any given video, to provide a variety of language and opinions. Altogether, we have collected 41,087 pieces of commentary. On average there are 7 pieces of commentary per minute of video, and typically the expert gives about 4 sentences of commentary every time they pause to react to the video. Table 4 shows the number of skills demonstrations with expert commentary per domain. Overall, we believe the commentary is a unique window into the skilled actions that (through language) surfaces many subtleties about the actions not evident to the untrained eye.

12.B. Narrate and Act

The participant surveys (Appendix 11) capture the *skill* of the participants, but they do not provide first-hand information regarding the *actions* being completed and *objects* they used. We capture these using “Narrate and Act”: a separate take recorded with the participant in which they narrate what they are doing and how they are doing it.

Participants were asked to complete the Narrate and Act take as if they are teaching someone else as part of an instructional video or how-to guide. In this way they would talk about what they were doing as they carried out an action. This is akin to the kind of narrations done by people creating how-to videos, though with less stylization and without any professional post-production editing. It is interesting to note that some participants did this naturally requiring little prompting. For some activities which were more intense, such as dancing or bouldering, we asked participants to instead narrate either just before or just after the action to reduce the difficulty of doing this live.

Compared to the third-person expert commentary (above), these narrations are first-person and delivered at the same time of the actions. Generally the commentary is richer in constructive feedback about the quality of the

activity, whereas the narrate-and-act narrations are interesting for their simultaneous nature and first-person analysis of what the participant is doing. The behavior in this extra take is expected to differ from that of the non-narrated tasks, in that it is likely that the participant will complete the scenario more slowly than normal to concentrate on explaining what they were doing. However, the narrate and act takes provided benefits in reducing annotation noise and they can potentially be used for multimodal learning as is currently explored in the literature with how-to video narrations [9, 86, 99, 100]. Table 3 (third column) and Figure 21 show examples.

12.C. Atomic Action Descriptions

Inspired by the “narrations” in Ego4D [44], our third language resource provides atomic action descriptions of the camera wearer and his or her environment, when relevant, as captured by the Ego-Exo4D videos. Compared to the expert commentaries and narrate-and-act narrations—which both emphasize the “why” and “how”—these atomic action descriptions are focused on the objective “what” of what is happening when, written in free-form text from the perspective of a third-party observer.

Annotation description We present each take to the annotators as a collaged video consisting of the Aria egocentric view, left and right grayscale SLAM, four or five fixed-position exocentric cameras, and single-track composite audio; for a subset of videos, a helmet-mounted GoPro view is also available. Annotators were asked to provide a play-by-play description of what happens, as seen across any of the views. Potential contents include actions by the Aria camera wearer, other individuals interacting with the camera wearer, and relevant environmental events. Each narration is atomic and time-anchored: as much as possible, each narration should roughly be limited to one verb and have a single associated timestamp, roughly within a second of its occurrence in the video. For consistency across narrations, the Aria camera wearer in each take is referred to as “C” in all narrations (e.g. “C picks up a wrench.”). Other individuals are referred to by other letters (e.g. “Man X kicks the soccer ball back to C.”); these letter labels are not necessarily consistent across takes, but refer to the same individual within a take. Many videos are narrated by two independent human annotators, and we make each narration set separately available. Table 3 (second column) and Figure 21 show examples. See Table 5 for atomic action descriptions summary statistics.

Visibility Because of the multi-view nature of the Ego-Exo4D capture rig, certain actions or events may not be visible across all camera feeds. While we hope Ego-Exo4D leads to increased attention toward multi-view learning,

Category	1x Coverage	2x Coverage	# of Descriptions	Descriptions Per Minute	Unique Nouns	Unique Verbs
Basketball	316	55	20301	53.887 (\pm 25.661)	192	137
Bike Repair	10	142	12365	24.601 (\pm 12.033)	459	270
Cooking	54	156	55968	25.243 (\pm 11.130)	1019	522
Dance	72	413	31119	29.146 (\pm 15.693)	470	444
Health	60	67	13385	22.848 (\pm 12.046)	404	254
Music	40	36	3750	10.638 (\pm 9.933)	180	119
Rock Climbing	77	79	6548	33.325 (\pm 10.897)	156	141
Soccer	0	36	5883	38.086 (\pm 21.256)	108	54
All	635	984	149421	31.497 (\pm 19.662)	1918	1071

Table 5. Atomic action descriptions per domain statistics.

many existing systems fundamentally assume a single view at a time; if a camera does not have a view of the narrated action or event, this may lead to a confusing learning signal, or pose an impossible ask for a model to infer. Thus, we also ask that the annotators answer two additional question per narration: 1) an indicator of whether the narration is visible from the egocentric camera, and 2) which (if any) of the static exocentric cameras provide the best view. If there are multiple equally good views, annotators are free to pick any. In particular, we found this *best exocentric view* helpful for other Ego-Exo4D annotation efforts: the narration visibility tags played a role in exocentric view selection for both the correspondences benchmark (Section 13.A.1) and expert commentary (above), and frame selection for hand and body pose (Section 13.D).

Relation to other Ego-Exo4D annotations While bearing some similarities to our other annotation workstreams, these atomic action descriptions exhibit some key differences:

- *Keysteps* (Section 13.B.1): While some atomic descriptions are similar to certain keystep names, narrations are free-form, without intentionally imposing any shared vocabulary. Atomic descriptions also aim to capture all actions and events, regardless of relevance as immediate steps towards the scenario’s goal, and cover all domains, not just procedural.
- *Expert Commentary* (Section 12.A): As atomic description annotators do not necessarily have the years of experience of the domain experts, the resulting text is often written from a layperson’s perspective, and focus on what is happening as opposed to providing critiques, analysis, or explanation. Atomic descriptions also tend to be shorter and denser than expert commentary.
- *Narrate and Act* (Section 12.B): While also a language resource, narrate and act’s descriptions are expressed from a first person perspective and can share intrinsic motivation behind actions. In contrast, atomic descriptions are written from the perspective of an outside observer.

12.D. Comparison of the Language Statistics

In Figure 24 we further emphasize the characteristics of each text corpus across three axes: total vocabulary size, average number of captions per video, and caption length. See caption for details.

Figure 25 shows word clouds per scenario and annotation type highlighting the differences in vocabulary and word frequency.

Contributions Statement

Kevin J Liang co-developed the expert commentary guidelines, interviewed and onboarded experts, helped test and suggest features for the narrator tool, and contributed to program management; he also developed the atomic action descriptions guidelines, helped coordinate annotations, and contributed to paper writing. Michael Wray contributed to the definition of the expert commentary guidelines; provided feedback for experts; co-developed the narrate-and-act guidelines and the pre-task/post-task questionnaires; and contributed to paper writing. Kristen Grauman proposed the expert commentary idea, co-developed the guidelines, interviewed and provided feedback to experts, and contributed to paper writing. Andrew Westbury implemented expert commentary, recruiting, mobilizing and managing our experts and workplan. Miguel Martin contributed to the atomic action descriptions annotation guidelines and produced the annotation files and associated tutorial code, and for expert commentary, he authored the initial version of the Narrator tool, transcribed the commentaries, and produced the annotation files and associated tutorial code.

Changan Chen contributed to the development of the narrator tool; provided feedback for experts; and contributed to paper writing. Siddhant Bansal contributed to the design of narrate-and-act, user questionnaires, object dictionaries, expert commentary cooking. Dima Damen proposed narrate-and-act data collection and user questionnaires and contributed to their design, and also contributed to expert commentary for cooking scenarios. Tiffany Davis provided significant program management support throughout expert commentary. Devansh Kukreja built the render flow to generate video collages for annotations.

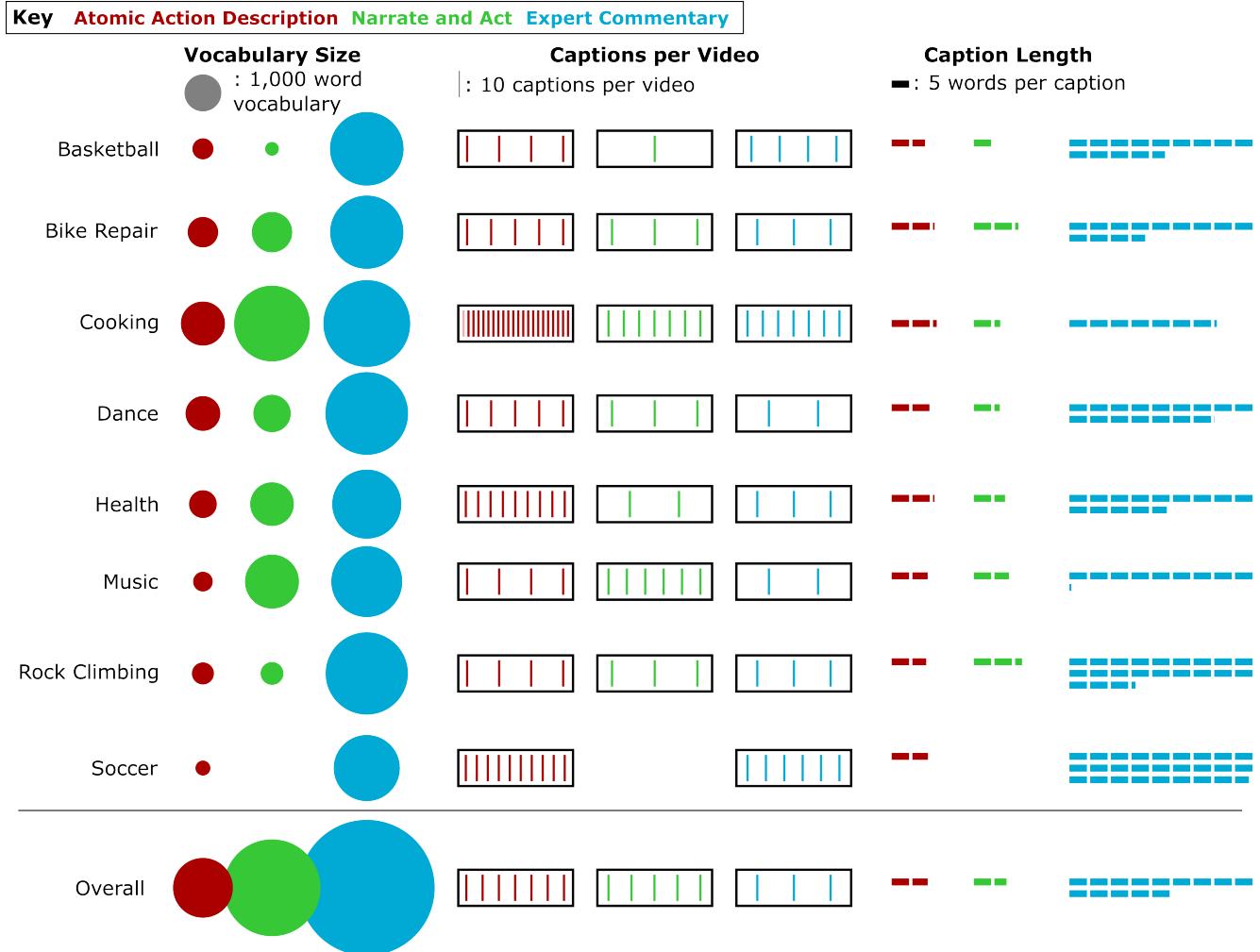


Figure 24. Comparisons between the vocabulary size (left) number of captions per video (center) and length of caption (right) for the atomic action description, narrate and act, and expert commentaries. Statistics are shown both per scenario and over the entire dataset. We see that the expert commentary tends to use a much larger vocabulary and more lengthy statements, since commentators are giving more elaborate statements of advice and explanation. The temporal density of the atomic action descriptions is greater than the other two forms, since the annotators are pausing to describe every single action of the camera wearer. Narrate-and-act comments use a vocabulary size in between the other two, reflecting the more free-form speech (compared to the written atomic actions) is used. Trends are mostly similar across scenarios, with the most noticeable differences being the temporal density; it is particularly high for both cooking and soccer. In the former, there are many procedural steps, whereas in the latter there are many instances of the drill being executed.

Domain resource people from our consortium were Dima Damen and Michael Wray [cooking], Kristen Grauman and Changan Chen [soccer], Gedas Bertasius [basketball], Kristen Grauman and Jianbo Shi [music], Andrew Westbury [health and bike repair], Kevin Liang [dance], Pablo Arbelaez and Maria Escobar [bouldering]

Ego-Exo4D’s panel of expert commentators is: **Soccer**: John Bello, Phillip O’Kennedy, Lee Bakewell, Radcliffe McDougald, Thomas Harris **Music**: James Peterson, Trevor Minton, Andrea LaPlante, Ethan Fallis, Alex Rogers, Jacqueline Burd **Health**: Jasmine Higa, Angela Liszewski, Kristin Blanset, Melissa Robinson, Sonya John-

son **Dance**: Rolanda Williams, Deanna Martinez, Enya-Kalia Jordan, Rachel Repinz, Yauri Dalencour, Kathryn Hightower **Cooking**: Mark Manigault, Mary Drennen, Tiffany Davis, Reginald Howell, Rosanne Field, Donnie Murphy, Kiet Duong, Laura de Vera, Keegan Taylor **Bouldering**: Daniel Ramos, Mike Kimmel, Roy Quanstrom, Christopher Deal, Carmen Acuna, Kelsey Hanson **Bike repair**: Cesar Pineda, Walker Wilkson, Frank Trotter, Cordell Bushey, Dimitri Elston, Sam Arsenault, Aaron Hill **Basketball**: Elizabeth Blose, Raven Benton, Joseph McCarron, Cornelius Gilleyen, Cecil Brown, Aaron Jones



Figure 25. Word clouds for each scenario and annotation type. The vocabulary for atomic action descriptions typically focuses on the person's hands and how they complete the actions (e.g. using left/right/hand) whereas narrate and act describe the high level goals/objects. The expert commentary has the largest variety of words, including specialist words for each scenario such as swab/solution for health and axle/valve for bike repair.

13. Benchmarks: Annotations and Baselines

In this section we provide details on all the benchmark task definitions, their annotations, and our baseline models and results. We cover ego-exo relation (Sec. 13.A), followed by ego-exo recognition (Sec. 13.B), proficiency estimation (Sec. 13.C), and ego body and hand pose (Sec. 13.D).

For each task, when formally defining the inputs and outputs, we also explicitly specify which inputs are *excluded* from use, if any. This is important information for future research done with Ego-Exo4D to ensure consistency of results. Table 6 provides summary annotation statistics for all tasks of Ego-Exo4D.

13.A. Ego-Exo Relation

The family of ego-exo relation tasks deals with relating the video content across the extreme ego-exo viewpoint changes, in the form of either object-level matching (correspondence) or generation of one view from the other (translation).

13.A.1 Ego-exo correspondence

Annotations We annotate pairs of temporally synchronized egocentric and exocentric videos with segmentation masks for selected object instances from six scenarios: *Cooking, Bike Repair, Health, Music, Basketball* and *Soccer*. We exclude *Bouldering* and *Dance* from this benchmark as they have limited diversity of objects. We focus on objects used by the camera-wearer at any point during the execution of the activity and that are visible in both views for at least some frames of the sequence. These masks allow us to define object-level correspondence between the views.

We used a multi-stage annotation process for annotating paired ego-exo videos:

- *Stage 0: Object Enumeration.* Annotator marks each object that is active at some point of the egocentric video with a bounding box in a frame where it is clearly visible and provides a free-form textual description.
- *Stage 1: Egocentric video annotation.* Annotator watches the egocentric video and is also shown (a) text and (b) a bounding box for one of the objects annotated in the previous stage. Annotator then marks a segmentation mask for that object in all the video frames where the object is visible. Segment Anything [67] is leveraged to generate segmentation masks efficiently using only point clicks.
- *Stage 2: Exocentric video annotation.* As shown in Figure 26, the annotator watches a temporally synchronized exocentric video and is also provided with the (a) text and (b) several ego segmentation masks of this object. Annotator then marks a segmentation mask for this object in all the exo video frames, whenever the object is visible.

What are the objects of interest? We focus on objects that are *active* at some point during the execution of the activity. These objects are not only interesting because they are essential to the activity, but they are also challenging to track, since they are moving/changing state. In particular, our annotation guidelines requested annotators to list (a) objects that the camera-wearer interacts with through their body or tools; (b) other objects that are relevant to the activity (e.g., supporting surfaces like kitchen top); and (c) body parts (hands and legs). Note that every time an object changes visual state (adopting the Point-of-No-Return definition from [44]), it is marked as a new object (e.g.,

Benchmark	Annotation Type	Number of Takes	Annotation Details
Ego-exo relation	Manual	907	2367 objects 282K ego masks 365K exo masks
Ego-(exo) keystep recognition	Manual	790	17 activities, 689 keysteps 20K ego segments (68 hours) 106K ego+exo segments (352 hours)
Procedure understanding	Manual	374	6 activities, 186 keysteps 5.4K segments (18 hours)
Ego-exo proficiency estimation	Semi-automatic	1419 (demonstrator)	1419 proficiency scores
	Manual	697 (demonstration)	14K “good” segments 16K “tips” segments
Ego pose	Automatic	521 (body) 521 (hands)	1.2M (3D), 5M (2D) 1.2M (3D), 6M (2D)
	Manual	203 (body) 103 (hands)	57K (3D), 375K (2D) 25K (3D), 143K (2D)

Table 6. Summary of annotation statistics for the different benchmark tasks of Ego-Exo4D.

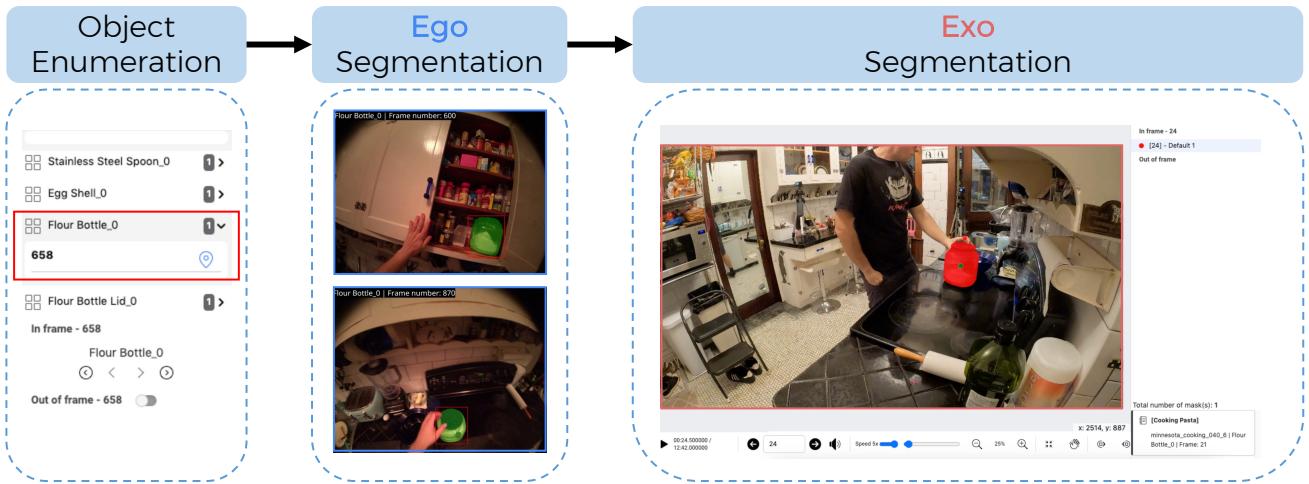


Figure 26. **Multi-stage annotation process for Ego-Exo Relation annotations.** After enumerating all active objects in the egocentric video, an object is selected and annotated with segmentation masks in all frames of the egocentric video. Then, annotators are given the exo video as well as the textual descriptions and sample egocentric segmentation masks for the object of interest, and mark segmentation masks for the specified object of interest in all the frames where it is visible.

Scenario	# Takes	# Objects	# Ego Masks	# Exo Masks
Basketball	393	602	21,818	31,143
Bike	80	186	15,227	19,956
Cooking	183	881	139,650	228,813
Health	127	523	69,327	77,212
Music	112	153	33,624	5,599
Soccer	12	22	2,411	2,475
Total	907	2,367	282,057	365,198

Table 7. **Relation annotation statistics.** We show statistics for each scenario including the number of takes, total number of objects annotated and the number of egocentric and exocentric segmentation masks.

annotators list *tomato* and *sliced tomato* as two distinct object instances).

Which objects to annotate with masks? For scenarios that involve few objects (Music, Basketball and Soccer), we annotated all object instances. Instead, for Cooking, Health and Bike Repair we sampled object instances based on their frequency of occurrence and their size, due to time and budget constraints. In particular, we binned each object annotated in the Object Enumeration stage into bins based on their frequency of occurrence across the dataset (high, low) and object size (small, large). We then uniformly sampled object instances from these bins while accounting for annotation time and budget and proceeded with segmentation mask annotations. We ignored all objects with area < 150 pixels. For Cooking, specifically, we also filtered out a few objects such as spices, mixtures and liquids, as

they tend to be too small to match in the exo view. Finally, we skipped exocentric mask annotations for objects that were visible in fewer than 10 frames of the egocentric video.

What frame rate to annotate at? We annotated segmentation masks at 1 frame per second, except for videos from the *Music scenario* which we annotated at 0.1 fps due to extremely long video durations.

In total, our annotation process yielded segmentation masks for 2,367 objects in 907 ego-exo video-pairs. Approximately 1.2 million frames were annotated resulting in a total of 647K segmentation masks. Table 7 shows a detailed breakdown per scenario.

Formal task definition Given a pair of time-synchronized egocentric and exocentric videos, as well as a query object track in one of the views, the goal is to output the corresponding mask for the same object instance in the other view for all frames where the object is visible in both views. This task is especially challenging in our dataset, since we have to handle long videos with an average length of 3 minutes, as well as very small objects with areas of only a few pixels. Importantly, note that the input to the model *excludes* semantic labels or names for the objects, camera pose information relating the two views, and IMU or active range sensor measurements. We do not use such information as we want to encourage the development of methods for open-world correspondence, not relying on predefined sets of objects or inputs that require non-consumer camera devices.

Metrics We adopt the following metrics in our evaluation:

1. *Location Error* (LE), which we define as the normalized distance between the centroids of the predicted and ground-truth masks.
 2. *Intersection Over Union* (IoU) between the predicted and ground-truth masks.
 3. *Contour Accuracy* (CA) [113], which measures mask shape similarity after translation is applied to register the centroids of the predicted and ground-truth masks.
 4. *Visibility Accuracy* (VA) [15], which evaluates the ability of the method to estimate the visibility of the object in the target view, as in practice it may often be occluded or outside the field of view. We measure this performance using balanced accuracy. Note that, in contrast to the previous metrics that compare segmentation masks at frames where the object is visible in both views, this metric is computed based on all frames with query masks.

Baselines Finding object mask correspondences across pairs of videos is an under-explored area in video understanding. Therefore, we investigate two diverse baseline approaches for our ego-exo correspondence task: (a) a *spatial model* that tackles the correspondence problem independently at each time point, and (b) a *spatio-temporal model* that takes into account the history of predicted correspondences.

Spatial baseline model. This model receives as inputs an egocentric frame, the associated exocentric frame, and a query object segmentation mask in one of the views. It then outputs the mask in the other view (if the object is visible in that view). It can be thought of as a generalization of query-point correspondence approaches proposed for sparse image correspondence [62]. We implement this baseline in the form of a Transformer-based image correspondence model, *XSegTx* (Cross View Segmentation Transformer), which extends *SegSwap* [138], a method originally proposed for image co-segmentation, i.e., for segmenting common objects in a pair of images. To adapt the architecture of *SegSwap* for our correspondence problem, we additionally condition the model on the segmentation mask of the object of interest by feeding the query mask as a third input to the model. In particular, we first pass the egocentric frame, the exocentric frame and the query mask (as a binary mask) through a visual backbone network. We then flatten the resulting features into three sequences and pass them through the cross-image transformer with alternating self-attention and cross-attention layers. We first use the query mask features to attend to the features in the query view which are then used to cross-attend over features from the target view. This allows the model to reason over features from both views conditioned on the input mask. The resulting sequences for

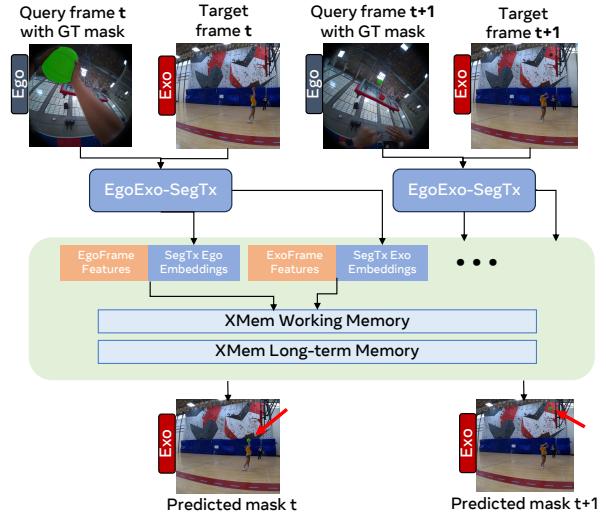


Figure 27. Overview of our spatio-temporal XView-XMem baseline model for the correspondence task.

both views are “unflattened” and passed through a decoder to predict object segmentation masks in both views. We also pass the target view features through a classification head to classify if the query object is visible in the target view.

We train the model to perform mask prediction using a point-wise binary cross-entropy loss and a dice loss over the predicted and ground truth masks. We use only pairs of frames where the object of interest is visible on both views and apply the losses on predicted masks in both the views. During inference, we only consider the mask predicted in the target view and discard the predicted mask in the query view. We train the head performing Visibility classification using a binary cross-entropy loss on all the frames of the sequence.

Spatio-temporal baseline model. The spatio-temporal model receives as input the pair of ego-exo video clips as well as an object segmentation track in one of the views, and outputs segmentation masks in the other view for the frames that the object is visible in both views. It can be thought of as performing generalized tracking across views. We build our baseline model on top of XMEm [22], a model originally proposed for tracking a specific target object given its segmentation mask in the first frame. In particular, our baseline model, called *XView-XMem*, adapts XMEm to track the object across different views given ground-truth segmentation masks for one of the views in each frame. To encourage the model to learn associations of the objects between egocentric and exocentric views, we train XView-XMem to track the object in a sequence of interleaved frames of egocentric and exocentric views, i.e., each egocentric frame is followed by an exocentric frame and vice versa, as shown

in Figure 27.

To mitigate track drift (within and across views), we also explore feeding the XSegTx embeddings to the XMem working memory. Since these embeddings are trained to guide the mask decoder at each frame independently, they capture rich information about the object of interest. The extracted image features from the ResNet in XMem are fused with the encoded embeddings from multiple layers of SA (self-attention) and CA (cross-attention) layers of XSegTx. They are then projected into keys and stored in memory for tracking.

Implementation details For our spatial baseline model, we downsample the images to 480x480 resolution for all the views while using padding to keep the original aspect ratio of the images. For the image backbone we use the same ResNet50 [48] checkpoint as SegSwap and freeze its weights during training. Our cross-image transformer architecture also follows [138]. We use a batch size of 32 and Adam [66] as our optimizer with a learning rate of 0.0002 which decays to 0.0001 after 50,000 iterations. We run all our experiments on a single Nvidia RTX A6000 GPU for 100,000 iterations.

For our spatio-temporal baseline model, we use the same visual backbone (ResNet50 [48]) and architecture as XMem [22]. Our only modification is in the information that gets inserted in the working memory at each frame. We first extract features from both ResNet and XSegTx for both both query and target frames. The corresponding features are then concatenated and projected to the original feature dimension through simple 2D convolution. We train on sequences of 8 interleaved ego and exo frames. The model is trained using AdamW as our optimizer with a learning rate of 0.00001 for 50,000 iterations and weight decay 0.05. The batch size is 8 clip pairs. We initialize our model with the original pretrained XMem, and keep both the ResNet backbone as well as our finetuned XSegTx models frozen. Note that we do not apply any data augmentations.

Data We use 812 takes from the Ego-Exo dataset to train and evaluate models for this benchmark. In particular, we use the common split shared across benchmarks, with 584 takes for training, 90 takes for validation and 138 takes for testing. We extract pairs of images between egocentric and exocentric views which have corresponding object masks annotated for training. This gives us a total of about 154k pairs for training.

Results We benchmark our XSegTx and XView-XMem baseline models on the test set in Table 8. We experiment with two settings: providing the ground-truth object track in the exo view (exo query mask) and predicting it in the ego view, and vice versa.

First, we observe that exploiting temporal cues helps with tackling the object correspondence task as shown by the significant increase in performance achieved by the spatio-temporal baselines (ST type) compared to the spatial ones (for example, IoU improves from 10.7% to 21.4% in the Ego→Exo setting.).

Second, we can see a big difference in performance between the Ego→Exo and Exo→Ego settings for all the baselines. In particular, models perform worse when the sequence of query masks is provided for the egocentric video and the model needs to predict query masks in exocentric video. This might be due to the heavy occlusion and very small size of objects in the exocentric views, making segmentation very challenging. While predicting a very tiny mask in the exo view can be very difficult, models can reason about the type and rough location of the object from a tiny mask in the exo view and thus accurately detect and segment it in the ego view, where it is much larger.

However, still all our baselines achieve a performance of $\leq 47\%$ IoU showing the challenging nature of the task and the dataset. We note that our dataset includes a great degree of object shape variation and high number of very small objects which are very difficult to model.

We break down our results across different activities in Fig. 28. We note that some activities are generally easier to model (e.g., basketball, soccer) because of limited variation in object shape and appearance whereas some activities (e.g., cooking and bike repair) are much harder to model due to larger diversity in appearance, shape and size of the objects across views. We also explicitly evaluate our baselines on their ability to predict masks for very small objects. To do so, we split our validation set based on the predicted object size in proportion to pixels in the image. We see that, all our baselines struggle on very small objects and perform increasingly well on larger object sizes. We also show some qualitative results in Fig. 30. As we can see, the spatial baseline (XSegTx) struggles to track the same object throughout the video. For example, in the bottom example, XSegTx alternates between predicting one and two object masks whereas the spatiotemporal baseline (XView-XMem) reliably tracks a single object throughout the sequence, showing the importance of exploiting temporal cues in the data.

Contributions Statement Manolis Savva co-led the correspondence benchmark and contributed to the task definition, the annotation guidelines, the baseline design, and paper writing. Effrosyni Mavroudi co-led the correspondence benchmark and contributed to the task definition, the annotation guidelines, the baseline design, and paper writing. Lorenzo Torresani contributed to the task definition, and to editing this section. Sanjay Haresh developed the spatial baselines and contributed to data analysis, experi-

Query Mask	Method	Type	Vis. Acc. \uparrow	IoU \uparrow	Location Error \downarrow	Contour Acc. \uparrow
Ego	XSegTx (random weights)	S	50.00	0.41	0.133	0.111
Ego	XSegTx	S	53.86	10.65	<u>0.139</u>	<u>0.295</u>
Ego	XMem (w/o finetuning)	ST	35.37	5.74	0.174	0.077
Ego	XView-XMem (w/ finetuning)	ST	96.90	<u>13.50</u>	0.179	0.208
Ego	XView-XMem (+ XSegTx)	ST	<u>89.20</u>	21.40	0.113	0.314
Exo	XSegTx (random weights)	S	50.00	0.71	0.118	0.369
Exo	XSegTx	S	50.00	15.66	0.143	0.201
Exo	XMem (w/o finetuning)	ST	59.67	<u>36.11</u>	<u>0.122</u>	<u>0.367</u>
Exo	XView-XMem (w/ finetuning)	ST	97.81	46.59	0.098	0.506
Exo	XView-XMem (+ XSegTx)	ST	<u>91.60</u>	35.29	0.150	0.360

Table 8. Baseline evaluation on the correspondence benchmark. Best results are reported in bold, second best results are underlined.

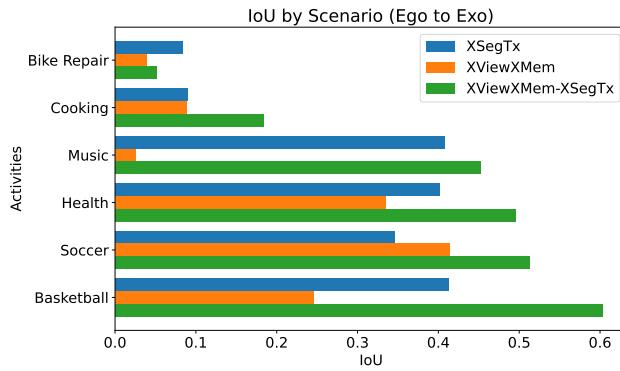


Figure 28. Performance of both baselines per activity scenario.

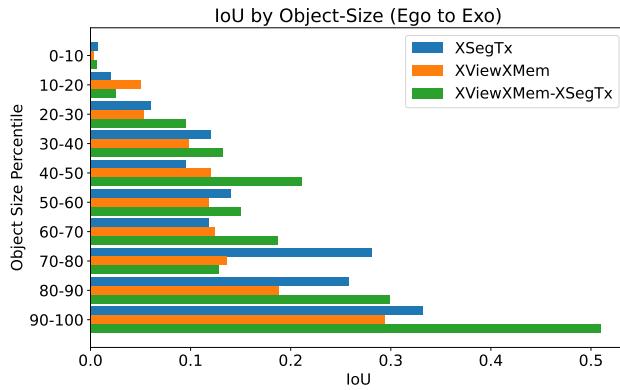


Figure 29. Correspondence evaluated across different object sizes in the target (exo) view. The object sizes range from $7e^{-6}\%$ to 11% pixels in the target view.

mental results, and paper writing. Yongsen Mao developed the spatiotemporal baselines and contributed to data analysis, experimental results, and paper writing. Suyog Jain formulated the annotation pipeline, developed annotation tools and contributed to annotation guidelines and paper writing. Santhosh Ramakrishnan contributed to the annotation

guidelines and the formulation of the annotation pipeline. Xitong Yang contributed to the annotation guidelines and the task definition. We would like to acknowledge Hanxiao Jiang for helpful discussions and preliminary ideas on baseline implementation. Devansh Kukreja built the render flow to generate frame-aligned videos of each camera for each take as model input.

13.A.2 Ego-Exo translation

Annotations Translation uses the same annotations as the correspondence task.

Formal task definition The translation benchmark focuses on generating information in the egocentric view given the exocentric view. The task is separated into two subtasks (see Fig 4): 1) *Ego Track Prediction* requires predicting the segmentation mask of an object in the unobserved ego frames given the object masks in an observed exo clip; 2) *Ego Clip Generation* entails generating RGB image values within the given masked ego view based on the exo view. For both subtasks, the input exo clip consists of 5 frames evenly sampled from a time span of 5 seconds.

Note that we restrict the input to include only the exo view and the object masks in order to promote the design of methods that can translate arbitrary third-person video into an egocentric one. Thus, the input *excludes* depth maps, 3D point clouds, IMU, or SLAM, which would simplify the task at the expense of general applicability, since these signals are typically not available for in-the-wild video. The only exception is a variant of the task where the ego camera pose for all frames of the clips is given as input. We consider this formulation in order to estimate a sort of “upper bound” on translation performance under the unrealistic assumption of known ego-exo camera relation.

Metrics We adopt a diverse set of metrics to assess the different aspects of the generated translation. As for the



Figure 30. Qualitative results for the different correspondence baselines.

task of correspondence, we use *Visibility Accuracy* (VA) to evaluate the ability of the method to predict the visibility of the target object in the ego view but this time given only exo frames as input. Furthermore, we adopt the following metrics defined for correspondence to gauge the performance of *Ego Track Prediction*: 1) *Location Error* (LA) 2) *Intersection Over Union* (IoU) and 3) *Contour Accuracy* (CA) [113]. The IoU and CA are calculated after registering the centroids of the predicted mask and the ground-truth ego mask, in order to gauge mask prediction independent of location error. To evaluate *Ego Clip Generation* we use two popular image quality metrics (SSIM, PSNR [51]) and three perceptual metrics (DISTS [30], LPIPS [187] and CLIP similarity [120]).

Baselines For track prediction, we implement the GAN-based method pix2pix [58] and the NeRF-based method GNT [150]. For clip generation, we employ the GAN-based method pix2pix [58] and the diffusion model DiT [112]. It is worth noting that, as discussed below, we introduce specific modifications to adapt these methods to our task requirements. All baselines utilize exo images and masks, with only the GNT model making use of the extra input of ego camera pose.

Ego Track Prediction involves generating segmentation masks for the egocentric view based on the exocentric video clip and the exocentric object masks. We consider the following two baselines for this task:

- **pix2pix-mask.** We modify the generator of pix2pix to have inputs and outputs of 4 channels. Specifically, the exo frame and the exo mask are concatenated as the inputs while the 4-channel outputs are ego frame (3 channels) and ego mask (1 channel). The ego frame is supervised with the losses used in pix2pix. We use the bootstrapped cross-entropy loss [125] and the dice loss [148] for mask prediction.
- **GNT-mask.** We adopt the Generalizable NeRF Transformer (GNT) [150] as another baseline leveraging the camera poses. In our adapted version, the image encoder takes a 4-channel image (exo frame and mask) as inputs to predict the ego frame and ego mask. Formally, during the training of our GNT-mask, for each point x and viewing direction unit vector $d \in \mathbb{R}^3$, the ray transformer f in GNT predicts two key attributes: RGB Color (c) and Object Existence Score (e), in which e signifies the probability of an object being present at point x . During rendering, the volumetric radiance field encoded by the ray transformer can then be rendered into a 2D image as well as a 2D object mask.

Ego Clip Generation requires producing pixel values representing the target object in the egocentric view. To achieve

Method	Ego Cam. Pose	Location Error \downarrow	Contour Acc. \uparrow	IoU \uparrow
pix2pix-mask	No	21.0	7.1	8.3
+multi-frame	No	21.9	7.0	7.0
GNT-mask	Yes	18.9	16.3	9.6

Table 9. Evaluation of translation baselines for the subtask of ego track prediction.

this, we leverage 6 different input images for each frame: exo frame, exo mask, exo object crop, cropped exo mask, ego mask and cropped ego mask. The cropped exocentric and egocentric masks are generated by considering a bounding box to isolate the relevant portions of the exocentric and egocentric masks, respectively. The “exo object crop” refers to the RGB image obtained by cropping out the relevant region using the cropped exocentric mask. We resize these 6 images to the same size (256×256). We evaluate two baselines for this task:

- **DiT-pix.** We adopt the Transformer-based diffusion model DiT [112]. We predict the ego object crop by conditioning the DiT on the 6 input images in two manners. Initially, these six images are concatenated along the channel dimension and subsequently combined with the noisy ego object crop, forming the input to DiT. Additionally, two ResNet-50 architectures encode the six images into low-dimensional features, which are then incorporated into each layer of DiT via AdaLN [114].
- **pix2pix-pix.** We adopt pix2pix [58] for clip generation as well by concatenating the 6 images along the channel dimension as inputs to the pix2pix model.

All of the above-mentioned baselines perform image-to-image generation. We implement also clip-to-clip variants of these methods by taking multiple frames as inputs and predicting results for all frames jointly. For pix2pix, we achieve this by replacing the original 2D-Conv with 3D-Conv, and 2D-BatchNorm with 3D-BatchNorm. For DiT, we use space-time divided attention as in TimeSformer [13].

Results We employ the validation set for the purpose of selecting optimal checkpoints and hyper-parameters, which are subsequently evaluated on the test set.

In the context of Ego Track Prediction (Table 9), we gauge performance using *Visibility accuracy* (VA) to assess the ability of the model to estimate the object visibility in the ego view. Correctness is determined when two conditions are met: (1) the predicted mask is empty when the object is invisible in the ego view, and (2) the predicted mask is non-empty when the object is visible in the ego view. Both pix2pix-mask and GNT-mask fail to perform well in estimating the object visibility, achieving Visibility accuracy

Method	SSIM \uparrow	PSNR \uparrow	DISTS \downarrow	LPIPS \downarrow	CLIP \uparrow
pix2pix-pix	0.52	15.9	0.38	0.59	73.0
DiT-pix	0.55	15.5	0.32	0.49	83.0

Table 10. Evaluation of translation baselines for the subtask of ego clip generation.

barely above the 50% performance of random guess (52.1% for GNT-mask and 50.2% for pix2pix-mask). However, the ResNet-50 trained exclusively to attend to this binary classification achieves a VA of 75.6%. We assess mask quality by considering distance (Location Error) and similarity metrics (IoU and Contour Accuracy) between predicted and ground-truth masks after registration. The 3D-aware NeRF-based baseline, GNT-mask, outperforms the implicit baseline, pix2pix-mask, overall. However, it does so by exploiting the ego camera pose as additional input. It is noteworthy that both baselines perform poorly on this task, likely due to the inherent challenges in correctly predicting the location and shape of the target object in the ego view, probably due to the fact that it often has diminutive size in the exo view.

In the case of Ego Clip Generation (Table 10), the Diffusion model DiT-pix demonstrates superior performance across all metrics compared to the GAN-based pix2pix-pix. Qualitative results (Figure 31) illustrate that DiT-pix can generate highly photorealistic images, aligning closely with the ground-truth in most instances. However, there are occasional cases (the last 2 rows) where the shape of the object is accurately generated, but the texture deviates slightly.

We further verify the importance of each input in Figure 32. Without exo object crop as input, the model fails to correctly infer the color and texture of the target object in the ego view. This result is expected as the source objects often represent a very small region of the entire exo frame. Additionally, without the ego crop mask as input, the model predicts the orientation of the object incorrectly. These observations highlight the importance of the cropped inputs.

We can observe in Table 9 that multi-frame (i.e., clip-to-clip) prediction does not provide a quantitative advantage over frame-to-frame prediction. Yet, we noticed that the multi-frame variant often yields generations that are more consistent across frames, even for frames where the exo view is heavily occluded, as can be seen in Figure 33. This is reasonable as a clip-level model can more effectively learn about the target object from multiple frames and fill-in information that is missing in individual exo frames.

In Table 11 we provide a break down of the results across different scenarios, using GNT-mask for track prediction and DiT-pix for clip generation. We can observe similar trends for the two subtasks: the methods achieve better results in basketball and soccer scenarios than in bike and cook scenarios, which is reasonable as the objects in bike and cook scenarios are more complex and diverse.

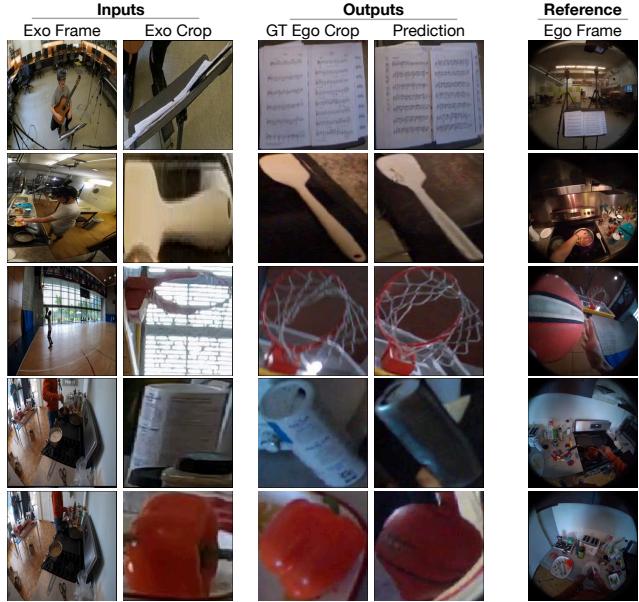


Figure 31. Qualitative ego clip generation by DiT-pix on the test set. The model takes 6 input images (exo frame, exo crop, exo mask, exo crop mask, ego mask, and ego crop mask). Note that only the exo frame and the exo crop images are included in this visualization. The ego frame, serving solely as a reference, does not constitute either an input or an output element.

Scenario	IoU (%) \uparrow	LPIPS \downarrow
Basketball	14.3	0.38
Soccer	13.8	0.46
Music	2.6	0.35
Health	6.3	0.42
Bike	6.0	0.52
Cook	9.7	0.51

Table 11. Breakdown of results per scenario for the subtasks of ego track prediction (IoU) and ego clip generation (LPIPS).

Contributions statement Lorenzo Torresani co-led the translation benchmark, developed the task formulation, advised the baseline development, and contributed to editing this section. Judy Hoffman co-led the translation benchmark and advised the baseline development. Feng Cheng led the baseline development and implemented the pix2pix and DiT models for track prediction and clip generation. Mi Luo implemented the GNT baseline and the evaluation pipeline for ego track prediction. Ziwei Zhao contributed the pix2pix baseline for multi-frame input and led the evaluation for ego clip generation. Huiyu Wang advised the baseline development and contributed to the task definition, the baseline design, and the metric selection and analysis.

13.B. Ego-(Exo) Keystep Recognition

This family of tasks centers around recognizing the keysteps

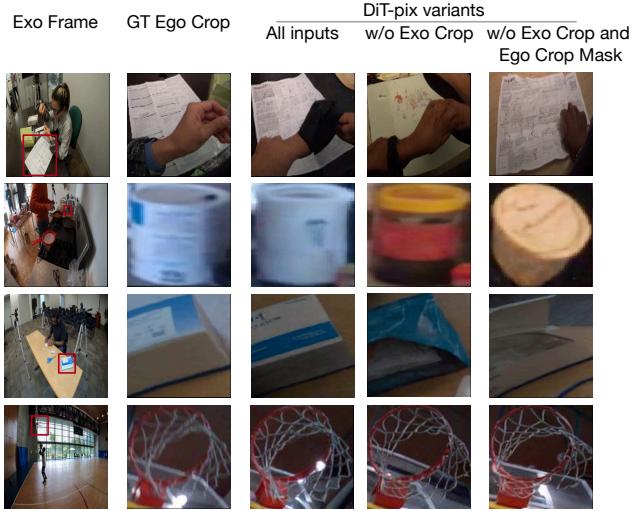


Figure 32. Qualitative demonstration of the importance of the different inputs given to DiT-pix. The exo crop image and ego crop mask are critical for good performance.

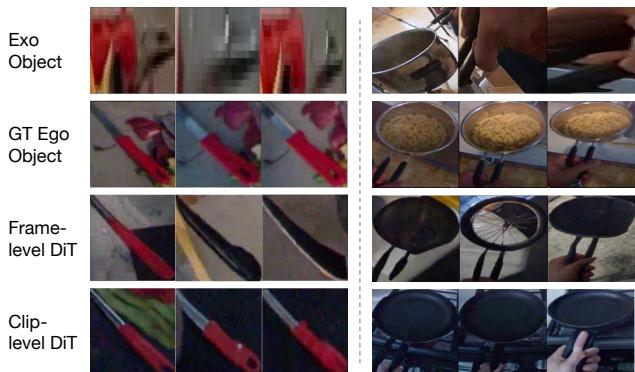


Figure 33. Comparison of ego clip generations using frame-to-frame vs clip-to-clip variants of DiT-pix. The clip-to-clip version of the model produces outputs that are more coherent across the frames of the clip, even for frames where the exo view is heavily occluded.

of a procedural activity and modeling their dependencies. Specifically, there are three tasks: fine-grained keystep recognition (Sec. 13.B.1), efficient multimodal keystep recognition (Sec. 13.B.2), and procedure understanding (Sec. 13.B.3). We refer to the family of tasks as “ego-(exo)” since exo may be available at the time of training but not inference.

13.B.1 Fine-grained keystep recognition

Annotations We annotate videos featuring any of the three procedural activities (i.e., cooking, bike repair, health) with temporal segments of *keysteps*, i.e., actions that contribute towards the completion of a procedural task. To ac-

curately model the hierarchical nature of the activities, we also develop a data-driven hierarchical keystep taxonomy concurrently with the annotation process.

Figure 34 shows the annotation user interface. We provide annotators a composite view of time-synchronized ego and exo videos. Each keystep annotation contains the start and end timestamps, a category label, a natural language description, and a flag indicating whether the keystep is essential or optional for task completion. Annotators interact with a search widget which displays keystep labels with their complete path within a hierarchical tree, e.g., Making cucumber & tomato salad > Prepare dressing > To a bowl or jar > Add salt.

As the activities performed by the camera wearers are unscripted, it is not possible to establish a comprehensive keystep taxonomy prior to annotation. To address this challenge, we designed an iterative, data-driven process for taxonomy development. We first initialize the taxonomy using various resources including recipes and instruction articles from the Web. This initial taxonomy captures keysteps that are generally expected in the activities, but it is assumed to be incomplete for the specific variations the camera wearers performed in the recordings. Subsequently, in each iteration, annotators receive the current taxonomy and are instructed to add new keysteps when they encounter actions not represented in it (see Figure 35). Any newly added keysteps are kept valid only for the duration of each annotation session and are not visible in other sessions. After a batch of videos have been annotated, we review the newly added keysteps to ensure their validity and update the taxonomy before repeating the process. We finalized the taxonomy after the third iteration, after which we re-annotated the entire set of videos with the final taxonomy for consistency.

Formal task definition We consider trimmed video clip classification as the keystep recognition task. At training time we are given a labeled collection \mathcal{D} of ego-exo video clips: $\mathcal{D} = \{(\mathcal{V}_{ego}^{(1)}, \mathcal{V}_{exo}^{(1)}, y^{(1)}), \dots, (\mathcal{V}_{ego}^{(N)}, \mathcal{V}_{exo}^{(N)}, y^{(N)})\}$ where $y^{(n)}$ denotes the keystep label of the n -th sample. The video clips are manually trimmed from long procedural videos to contain only the keysteps to recognize. At test time, given *just* the ego view of a trimmed clip \mathcal{V}_{ego} , the model must predict its keystep label y .

Classification of trimmed video clips is a problem formulation commonly adopted in action recognition benchmarks [43, 64, 145]. However, our task differs from action recognition in two fundamental aspects. First, it targets fine-grained keystep recognition rather than classification of coarse activities. We note that this adds significant complexity, since different keysteps of an activity often involve manipulating the same objects in the scene (e.g.,

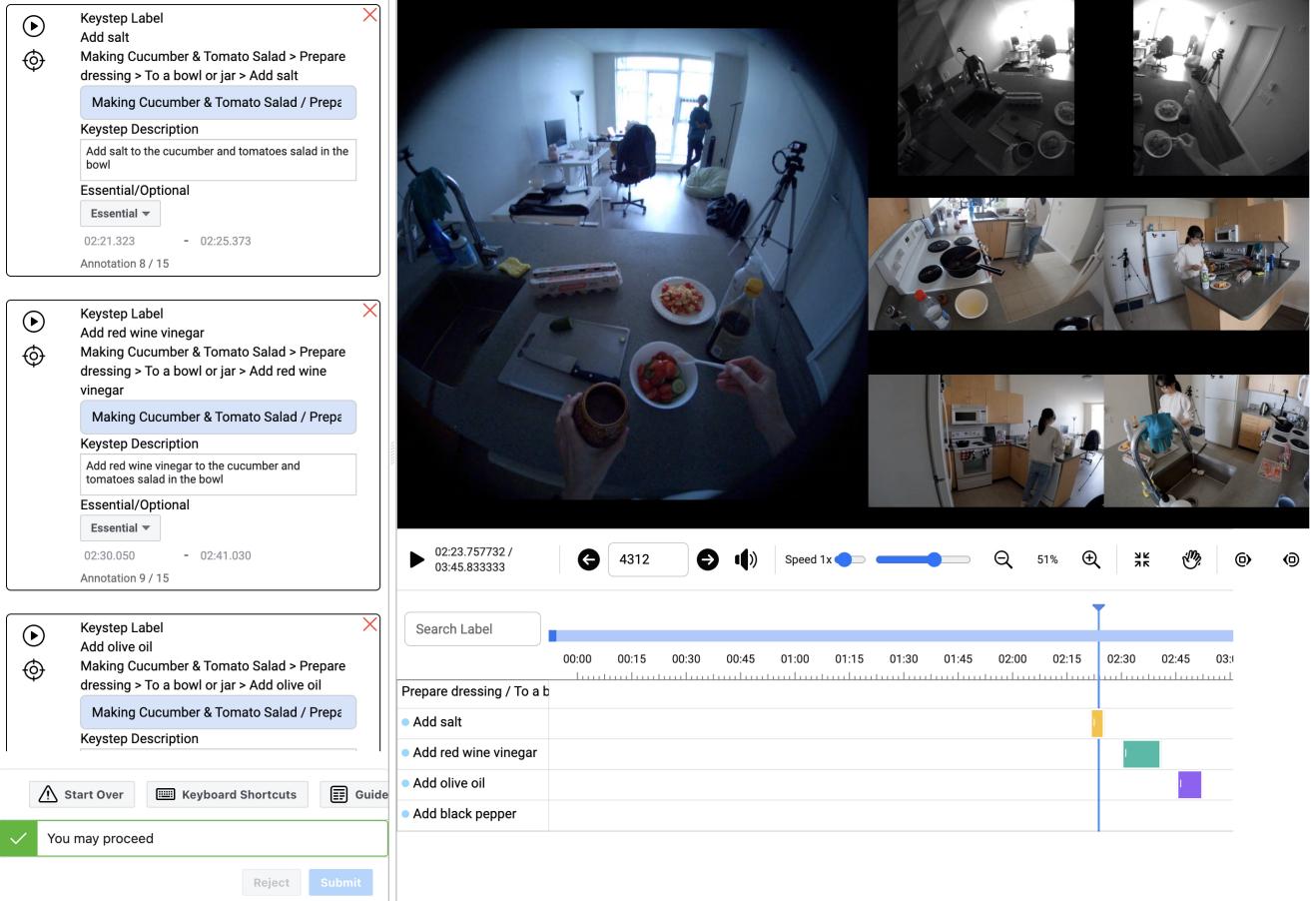


Figure 34. **The keystep annotation tool** shows a composite view of the time-synchronized ego-exo videos and the keystep time segment annotations. Each annotation consists of the start and end timestamps, a category label, a natural language description, and an essential/optional flag.

Scenario	Takes		Ego Keystep Segments		Ego + Exo Keystep Segments		Taxonomy	
	Count	Duration (total / avg [†])	Count (total / avg [†])	Duration (total / avg [‡])	Count (total / avg [†])	Duration (total / avg [‡])	Activity	Keystep
Cooking	337	51.50h / 551.79s	14,035 / 41.65	45.90h / 11.77s	73,877 / 219.22	243.03h / 11.77s	11	549
Bike repair	203	8.78h / 155.69s	1,601 / 7.89	7.77h / 17.47s	8,036 / 39.59	39.16h / 17.47s	4	84
Health	250	15.46h / 222.67s	4,770 / 19.08	14.07h / 10.62s	23,850 / 95.40	70.33h / 14.07h	2	58
Total	790	75.74 / 345.59s	20,406 / 25.83	67.73h / 11.95s	105,763 / 133.88	352.51h / 11.95s	17	689

Table 12. **Keystep annotation statistics.** We report the statistics by grouping our 17 activities into three scenarios: cooking (11), bike repair (4), and health (2). Statistics are listed for takes[†] and keystep segments[‡].

folding the bedsheet and smoothing out the bedsheet) and are consequently difficult to tell apart. Furthermore, different keysteps may be represented over largely different time spans (e.g., the average time span for “kneading dough” is 87.3 seconds, in stark contrast with “getting salt”, which averages at 3.6 seconds), thus requiring analysis at different levels of temporal granularity. The second key difference is the potential to leverage contextual cues available in exocentric videos during training to improve the prediction accuracy on egocentric videos. Note that at test time, the input to the model includes just the ego-view videos (RGB only).

Exo-view videos, activity and scenario names, narrations, audio and associated metadata such as eye gaze, 3D point clouds, camera pose, and IMU information are *excluded* as inputs for inference (although we encourage exploring their potential utility in training) as our ultimate goal is a vision-centric approach that performs egocentric keystep recognition.

Baselines To understand the best strategy for egocentric keystep recognition with paired ego-exo training data, we consider a diverse set of baselines approaches, includ-

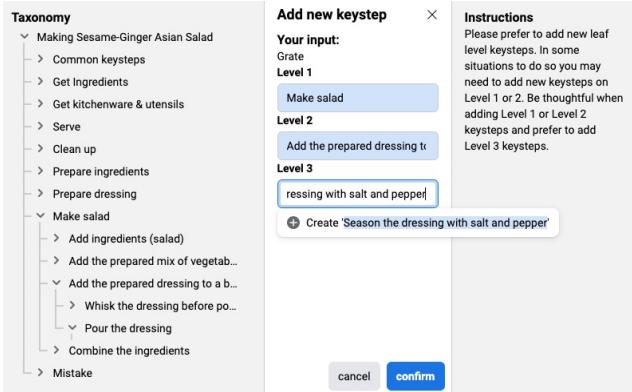


Figure 35. **Adding new keysteps to a taxonomy.** Annotators utilize a specialized widget to introduce new keysteps at any level within the existing taxonomy hierarchy.

ing methods for action classification, video representation learning, and ego-exo transfer.

- **Action classification.** As a prototypical example of this classic genre, we select a TimeSformer [13] model initialized with the checkpoint pretrained on the large-scale third-person action dataset Kinetics-600 [64] due to its strong performance in various video understanding tasks.
- **Video-language pretraining.** We adopt the EgoVLPv2 framework [119] and initialize the model with two backbones, one pretrained on the Ego4D dataset [44] (which contains only ego views) and the other that is further pretrained on EgoExo4D (which encompasses both ego and exo views), to exploit the variances in pretraining between these datasets.
- **View-invariant learning.** A two-stage training approach is employed. In the first stage, we utilize all available (ego, exo) video pairs in the dataset for training a view-invariant (VI) encoder. The training objective is a clip-level contrastive loss [106], aiming at identifying the synchronized (ego, exo) pairs as positive, and the non-synchronized pairs as negative. In the second stage, this pretrained model is further trained with a classification loss, aligning with the clip-level classification nature of the downstream task. Note that to align with the clip-level classification task, our contrastive loss operates at the clip-level, rather than at the frame-level as was done in view-invariant loss proposed in [136, 140].
- **Viewpoint distillation.** This also adopts a two-stage training approach. In the first stage, we train a multi-view teacher that takes both ego and exo views as input. In the second stage, a single-view ego student is trained, distilling knowledge [49] from the multi-view teacher to encapsulate information from both views.
- **Ego-Exo Transfer.** Here we follow the methodology proposed in Ego-Exo [79] which uses egocentric pseudo-labels to pre-train the network. We employ a masked au-

toencoder (MAE) [158] backbone, initialized from a Kinetics checkpoint, and the pseudo-labels provided from the Ego-Exo checkpoint to fine-tune with two auxiliary heads (Object-Score and Interaction-Map). We then further finetune the model with a classification loss for fine-grained keystone recognition.

For the first two baselines (which utilize pretrained checkpoints from well-established benchmarks), two training settings are further examined: one using only the ego view for classification loss and the other utilizing both ego and exo view videos, with the training objective being the sum of ego view and exo view classification losses.

Implementation details We use clips of size $8 \times 224 \times 224$, with frames sampled at a rate of 1/32 for all baselines except for EgoVLPv2 (where we adhere to its pretraining scheme and sample 4 frames). The patch size is 16×16 . For training, we resize the shorter side of the frame to a random value within the range of [256, 320], followed by randomly sampling a 224×224 region from the resized video. For evaluation, we sample a single temporal clip in the middle of the video, scale down the shorter spatial side of the video to 224 pixels and select 3 spatial crops (top-left, center, bottom-right) from the temporal clip to cover a larger spatial extent within the clip. The final prediction is derived by averaging the scores obtained for these 3 crops. We train our model for a total of 100 epochs on 4 NVIDIA V100 GPUs with a batch size of 32. The model checkpoint yielding the best performance on the validation set is selected and evaluated on the test set.

Data The keysteps in our dataset exhibit a very long-tailed distribution. To address this challenge, we set a cutoff threshold at 20 samples per keystone, limiting our analysis to 238 unique keysteps as shown in Figure 36. For simplicity, we consider only the leaf node keysteps in the hierarchy. Exploring the hierarchical structure including parent nodes is a promising direction but we leave this as future work. In all, the dataset for keystone recognition comprises 76,057 segments, with an average duration of 11.95 seconds each. Specifically, the training set contains 49,816 segments, of which 9,570 are from the ego view and the rest from the exo view. The validation set consists of 10,823 segments, including 2,127 ego-view segments, and the test set has 15,418 segments with 2,940 in the ego view.

Results Table 13 reports the Top-1 accuracy for ego classification on both validation and test sets. Among all the baselines, the TimeSFormer model pre-trained on Kinetics-600 emerges as the top performer, achieving a test accuracy of 38.27%. It is closely followed by the Viewpoint Distillation and the Ego-Exo Transfer MAE, which attain test

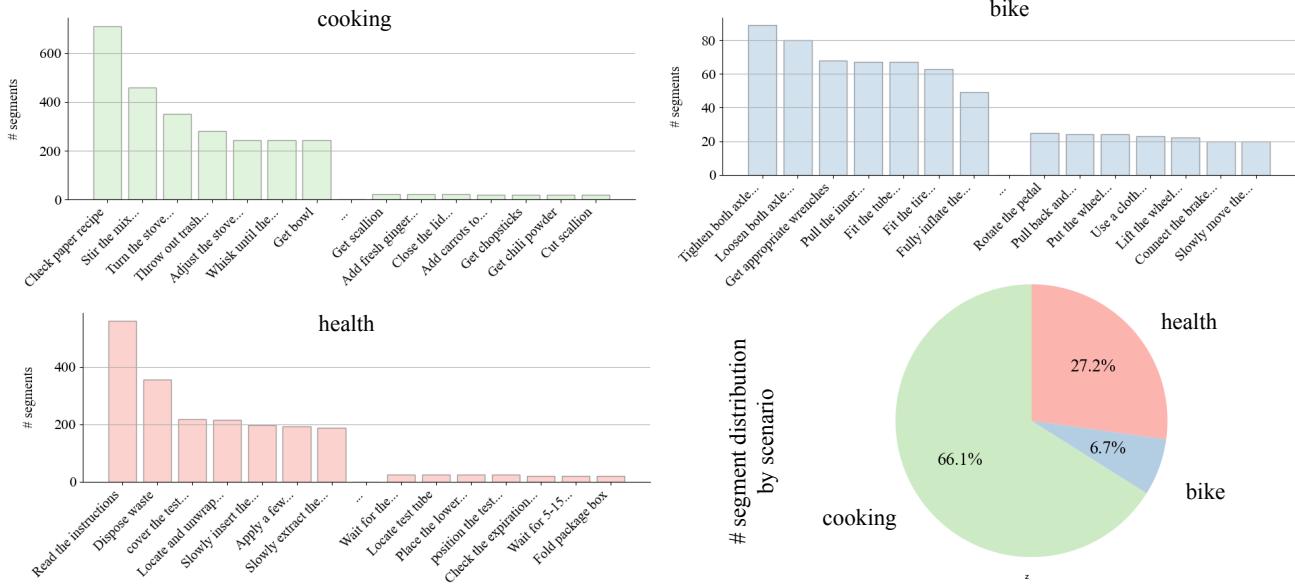


Figure 36. Keystep distribution in our dataset for each procedural scenario: cooking, bike repair, and health.

Method	Train data	Ego Accuracy (%)	
		Val	Test
TimeSFormer [13] (K600)	ego	44.66	38.27
TimeSFormer [13] (K600)	ego,exo	41.61	34.66
EgoVLPv2 [119] (Ego4D)	ego	41.70	34.90
EgoVLPv2 [119] (Ego4D)	ego,exo	42.78	36.26
EgoVLPv2 [119] (EgoExo)	ego	42.78	36.26
EgoVLPv2 [119] (EgoExo)	ego,exo	42.69	36.05
VI Encoder [106] (EgoExo)	ego,exo	43.31	36.46
Viewpoint Distillation [49]	ego,exo	43.91	<u>37.99</u>
Ego-Exo Transfer MAE [79]	ego,exo	<u>44.10</u>	37.55

Table 13. Top-1 accuracy of keystep recognition. The pre-training dataset is denoted in parentheses. VI is short for view-invariant.

accuracies of 37.99% and 37.55% respectively. The VI Encoder trails with a performance of 36.46%. We hope that these baselines provide insight to future work on how to best utilize exo-view videos available during training. Additionally, we note that the MAE, when trained without Ego-Exo signals, recorded a test accuracy of 36.83%. From all these results, it appears that additional exo-view videos do not substantially enhance the performance of egocentric keystep recognition. We identify two potential reasons.

First of all, as the exo cameras are typically positioned further away from the subject, the exo views may not capture the activity with sufficient details, especially if the activity involves manipulating small objects in front of the body (such as in cooking). To confirm this hypothesis, we analyze per-scenario performance and summarize the results in Figure 37. We compare the two training regimes

of EgoVLPv2 (Ego4D): one with ego-view videos ('ego-only') and the other with videos from both ego and exo views ('ego+exo'), reporting the final ego-view test accuracy. Additionally, we conduct an experiment using only the exo-view videos for both training and testing ('exo only'). The findings suggest that exo views can facilitate improved performance in the health scenario. A plausible explanation might be that the keysteps within the health scenario necessitate a more global understanding of the human body to make accurate predictions, unlike cooking, which benefits from the egocentric view to clearly observe the objects being manipulated (e.g., steps like getting turmeric powder). To further analyze this, we conduct a detailed analysis of per-step performance in Figure 38, comparing training with ego-view videos and exo-view videos. We can observe that exo views show performance advantages over ego views in several steps, predominantly in the health scenario, whereas ego views exhibit superior performance in steps associated with the cooking scenario. These results further support our hypothesis.

The second possible reason for the lack of gain is that our current baselines may fall short in learning view-invariance effectively, underscoring a significant opportunity for future work. There remains a substantial scope to explore methods to bridge the ego-exo view gap, and devise strategies to leverage useful signals from exo videos to augment ego accuracy. In addition, it is important to note that different backbones respond differently to the addition of exo-view videos during training. Specifically, while the TimeSFormer (K600) exhibits a slight degradation when

the exo classification loss is integrated into the objective, EgoVLPv2 (Ego4D) benefits from the introduction of exo-view videos. Overall, we posit that the endeavor to enhance view-invariant learning and to more effectively harness the complementary information from exo views for egocentric keystep recognition remains an open avenue. This encourages further investigation and innovation in this domain.

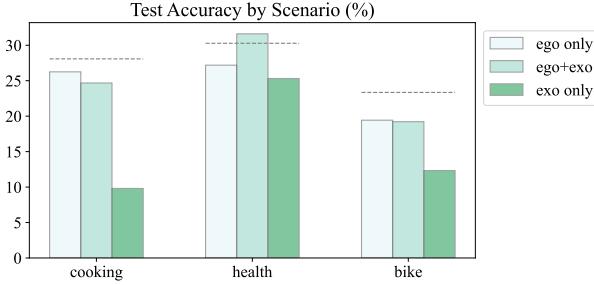


Figure 37. Keystep recognition evaluated per scenario. ‘ego only’ and ‘ego+exo’ denote the use of EgoVLPv2 (Ego4D) with ego-view videos and both views during training, respectively; note that for these two variants the final ego-view test accuracy is reported. ‘Exo only’ indicates a TimeSFormer model pretrained exclusively on exo-view videos, with the final exo-view test accuracy reported. Dotted line denotes the best performing model in Table 13, i.e., TimeSFormer (K600) trained on ego-view videos.

Contribution statement Tushar Nagarajan co-led the keystep recognition benchmark, co-developed the task formulation, and advised the baseline development. David Crandall co-led the keystep recognition benchmark and contributed to the task formulation. Yale Song co-led the keystep recognition benchmark and led the keystep annotation effort, including design of annotation guidelines, taxonomy development and coordination of annotation workflows; he also contributed to the task formulation, advised baseline design, and facilitated the delivery of EgoVLPv2 pretrained backbone. Triantafyllos Afouras contributed to the taxonomy definition, managed the labeling effort, and developed software for post-processing the annotations. Zihui Xue led the baseline development effort, implemented the TimeSFormer, EgoVLP, VI Encoder and Viewpoint distillation baselines, and performed analysis of results. Eugene Byrne contributed to the taxonomy development and to the Ego-Exo Transfer baseline implementation and analysis. Avijit Dasgupta contributed to the annotation and taxonomy development, and to the early stage of baseline design. Miguel Martin contributed to the annotation and the taxonomy development. Shraman Pramanick contributed the EgoVLPv2 pretrained backbone. Yifei Huang contributed to the early stages of task definition and baseline design. Devansh Kukreja built the render flow to gener-

ate frame-aligned videos of each camera for each take as model input, and to produce video collages for annotations. Lorenzo Torresani contributed to editing this section. Kristen Grauman contributed to the task formulation.

13.B.2 Energy-efficient multimodal keystep recognition

Annotations and data This task uses the same egocentric videos and annotations as keystep recognition. However, in addition to the raw RGB video, it uses the audio stream (and potentially other sensors) as another sensor modality.

Formal task definition In this task, the goal is to perform *online* classification of keysteps in a streaming egocentric multi-modal video, within an energy budget. We consider an ego video \mathcal{V}_{ego} of arbitrary length T comprising a stream of K different sensory modalities (e.g., RGB images, audio, IMU, etc.). At each time step t , where $1 \leq t \leq T$, the video consists of samples for each available modality, such that $\mathcal{V}_{ego}^t = \{S_1^t, \dots, S_K^t\}$, where S_j^t denotes the sample at time t for the j^{th} modality. Given \mathcal{V}_{ego} and an energy budget B , our task is to learn a model \mathcal{F} that maximizes the overall keystep recognition performance across the full video while also ensuring that the combined energy for sensing and running model inference does not exceed B . \mathcal{F} consists of a sensor triggering policy \mathcal{F}^P and a keystep prediction model \mathcal{F}^K . At every step t , the policy \mathcal{F}^P decides which sensor(s) to activate and sample from, in order to produce the model’s current observation O^t , such that $O^t \subseteq \{S_1^t, \dots, S_K^t\}$. Given O^t , the keystep predictor \mathcal{F}^K outputs its estimate of the ground truth keystep for the current step.

The energy budget accounts for the cost of operations in each model forward pass, the cost of moving intermediate activations in and out of memory and the cost of the continuous operation of sensors, each having different cost profiles (e.g., IMU and audio sensors are relatively cheaper to operate than camera sensors). Note that the sensor triggering policy may be static (e.g., sample video at 4 frames per second (fps), keep audio/IMU off; sample 1 fps video, keep IMU always on) or dynamic (e.g., depending on the audio, decide whether to trigger video capture). We keep our task definition general, allowing the challenge to admit a wide variety of recent approaches ranging from pure video-based efficient backbone architectures [38] to multi-modal triggering approaches and, naturally, a combination of them.

Note that at test time, the input to the model can only include current and past observations as our task is strictly an online recognition task. However, we encourage exploring other modalities than those considered in our experiments, e.g., IMU or camera poses inferred from video, audio, and

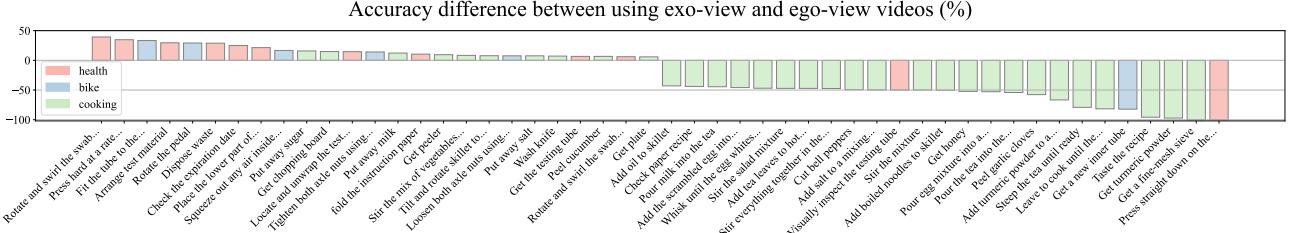


Figure 38. Keystep recognition evaluated per keystep label, comparing training with only ego-view videos versus exo-view videos. The accuracy delta (exo-ego) is displayed, where a positive value indicates better exo and a negative value indicates better ego performance.

IMU.

Measuring energy consumption Accurately measuring energy consumption of models is crucial for their use in AR/VR devices [2, 21]. The energy used comes from a complex interplay of sources including sensors, compute, communication, data processing, memory transfer (SRAM and DRAM), and leakage – many of which are typically ignored when building *efficient* computer vision models, despite their large energy consumption (e.g., memory transfer accounts for over 50% of the total power [178]).

We consider three key factors when modeling energy consumption following prior work [149]. (1) Compute energy: the cost of each model forward pass as a function of the number of operations (MACs). (2) Memory transfer energy: the cost associated with memory read-write operations for storing intermediate activations and model outputs. (3) Sensor triggering energy: the cost associated with turning on / off and continuous operation of sensors (camera, audio, IMU). For a model that processes an observation O^t , the total energy consumed can then be formulated as.

$$E(O^t) = \alpha * C(O^t) + \beta * M(O^t) + \sum_{j=1 \dots K} \gamma_j * \mathbb{1}(S_j \in O^t), \quad (1)$$

where $C(O^t)$ corresponds to the total number of multiply-add operations computed during the forward pass (in MAC/s), $M(O^t)$ corresponds to the total memory transferred to/from DRAM (in MB/s), and $S_j \in O^t$ corresponds to whether the j -th sensor is active. Finally, α, β, γ_j are weighting factors that measure the contribution of each energy source. We select these weighting parameters to reflect real-world AR/VR hardware capabilities. Namely, $\alpha = 4.6$ pJ/MAC [28, 149]; $\beta = 80$ pJ/byte [52]; $\gamma_{rgb} = 15$ mW and $\gamma_{audio} = 0.5$ mW [88].

While physical measurements are required for truly accurate energy estimates, this formula provides a reasonable approximation to it and can serve as a target for the computer vision community towards energy-efficient models for real-world devices.

Energy profiler We adapt off-the-shelf profiler software built for PyTorch to compute the total multiply-accumulate operations (MACs) and memory transfer (MB) required to estimate total energy in Eqn. 1. The quantities are time-normalized — total energy consumption is expressed as power (mW). We describe each component of the profiler below.

- **Compute operations (MACs)** We use the native PyTorch FLOP counter to get the total FLOP count in the forward pass. We convert this to MACs (approximately 2 FLOPs = 1 MAC).
- **Memory transfer (bytes)** We consider GPUs as our processing device, and use the PyTorch memory profiler to get the list of all operations executed in the forward pass (`model.forward()` call) and their associated GPU memory usage. The total memory is the sum of the individual operation memory costs.
- **Sensor capture** For each modality, we measure the time for which it is active as the number of observations sampled containing the modality. We require that the sensors capture at least 1 second worth of samples (roughly 100 samples) as energy consumption is ill-defined for an *instantaneous capture*.

Metrics Following prior work [27], we evaluate online keystep detection performance using per-frame calibrated mean average precision (mcAP), which accounts for the imbalance in the keystep labels in our dataset. We measure energy consumption in mW as described above.

There is a natural trade-off between efficiency and better performance. Thus, we evaluate models in two tiers by setting a budget for the power consumption in each tier, namely 20 mW for the *high-efficiency* tier and 2.8W for the *high-performance* tier. We select the *high-efficiency* budget based on the energy consumption of current single-modality, efficient architectures (e.g., X3D-XS [38]) with an eye to the future where multi-modal models operate within it. For the *high-performance* tier, we set the budget to a value that permits the use of powerful transformer-based action recognition models like LaViLa [194]. Once

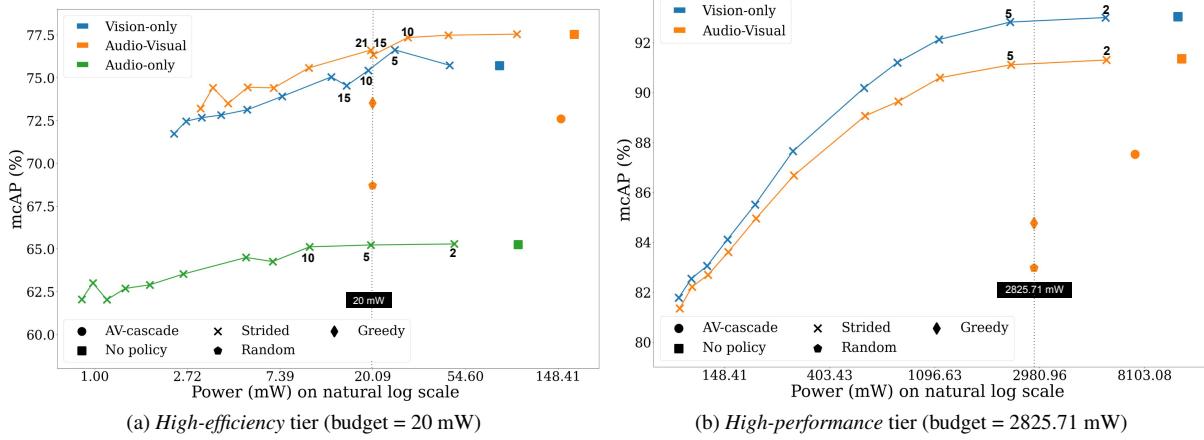


Figure 39. Keystep prediction performance (mcAP) vs. total power consumption with different prediction backbones and sampling policies for both *high-efficiency* (left) and *high-performance* (right) tiers. For the models using a fixed stride, we show their stride value in text if their total energy consumption is close to the budget.

a model runs out of budget, in our setup it uses its latest prediction for all future steps.

Experimental setup We instantiate the task by considering keystep prediction episodes where the multimodal samples arrive in a streaming fashion. As mentioned above, we use vision and audio as our task modalities, where vision comprises RGB frames that are streaming at 30 frames per second (fps), and the audio modality is made up of time-aligned single-channel chunks that are 0.4 seconds long and sampled at 16 kHz. However, the setup can be extended to include IMU, and potentially other sensors as well. We evaluate all models at the rate of 5 fps on a total of 109 test episodes of variable length, where the shortest episode is \sim 20 seconds, and the longest episode is \sim 25 minutes.

Baselines We provide a family of (less/more expensive) keystep prediction models for solving the task. Each model has a unimodal or audio-visual feature encoder followed by a keystep classification head.

- **X3D-XS [38].** This is a vision-only model comprising the X3D-XS feature encoder, which progressively expands the feature size and representational capacity of its layers, and later contracts them for achieving better performance-efficiency trade-off. This is the most lightweight model in our family of keystep predictors. The encoder has a depth factor of 2.2, and takes 4 RGB frames of size 160×160 sampled at 15 fps, as inputs.
- **LaViLa [194].** This is another vision-only model where the visual feature encoder is trained through CLIP-style video-language pre-training. To improve the feature quality over vanilla CLIP-style pre-training, this method augments the number of video-text pairs by leveraging pre-

trained large language models (LLM) to generate textual descriptions of un-annotated videos and rephrase existing narrations. In particular, we use the frozen TimeSformer [13]-Base (TSF-B) visual encoder pre-trained on the Ego4D dataset. To generate the feature for a target frame, the encoder samples 12 RGB frames of size 224×224 at 30 fps from a time window centered around the target frame and pads the samples with the boundary frames on both ends to create a 16-frame clip.

- **Light-ASDNet [80].** This is an audio-only model that represents audio as spectrograms and efficiently encodes them by splitting 2D convolutions into 1D convolutions along the spectrogram temporal dimension [80]. In our setup, the spectrograms are Kaldi [118]-compliant, and consist of 196 temporal windows and 160 Mel-frequency bins, respectively.

- **Audio-Visual Late Fusion (AV-LF).** This is an audio-visual model that does late fusion of visual features (encoded with X3D-XS or LaViLa) and audio features from Light-ASDNet by using linear layers.

To improve the energy efficiency of the aforementioned keystep predictors, we employ the following baseline policies for determining when to sample or skip each modality:

- **Fixed stride.** This is a policy that samples the input (video or audio) every s prediction steps. We evaluate different s values, where $s \in \{2, 5, 10, 15, 21, 43, 64, 86, 107, 129, 150\}$.
- **AV-LF + greedy.** This is a policy that greedily uses up the budget by sampling both audio and vision as early as possible, and uses the AV-LF backbone for keystep prediction.
- **AV-LF + random.** This is a policy that randomly samples or skips the audio and/or visual inputs until it runs out of budget, and uses the AV-LF backbone for prediction.

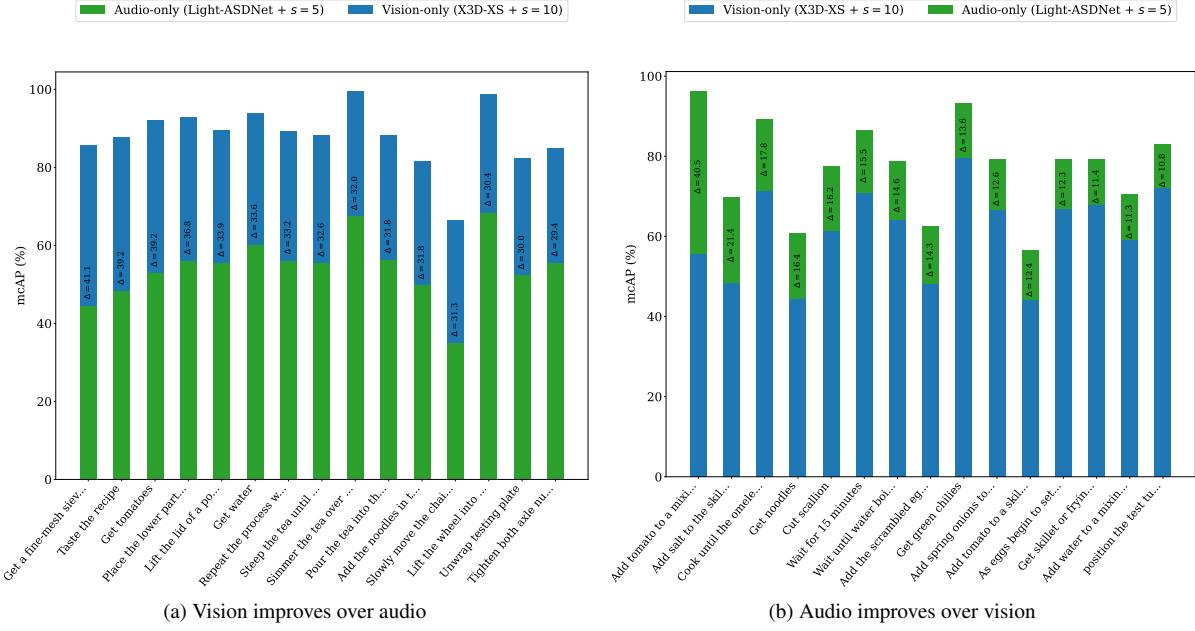


Figure 40. Improvement (left) or degradation (right) in keystep recognition performance per keystep label, when comparing the most efficient vision-only (X3D-XS [38] + $s = 10$) and audio-only (Light-ASDNet [80] + $s = 5$) models from the *high-efficiency* tier. The plots show the 15 keysteps where improvement or degradation are largest. Δ reports the amount of improvement/degradation.

Method	Modality	mcAP (%) ↑	Power (mW) ↓
Light-ASDNet [80] + $s = 5$	A	65.22	19.70
X3D-XS [38] + $s = 10$	V	75.42	19.12
AV-LF w/ X3D-XS + $s = 15$	AV	76.61	19.66

Table 14. Keystep prediction results for the *high-efficiency tier* (budget = 20 mW).

Method	Modality	mcAP (%) ↑	Power (mW) ↓
Lavila [194] + $s = 5$	V	92.82	2251.92
AV-LF w/ Lavila + $s = 5$	AV	91.11	2271.65

Table 15. Keystep prediction results for the *high-performance tier* (budget = 2.8W).

- **Audio-Visual (AV) Cascade.** This is a policy that initially uses the Light-ASDNet model to predict the keystep, and switches over to the LaViLa model (expecting a more accurate prediction from it) if the audio-based prediction confidence is below a certain threshold. We set the confidence threshold to 0.5 in our experiments.

Implementation details We train all keystep prediction models for 150 epochs using the cross entropy loss. We use the AdamW [92] optimizer with an initial learning rate of 10^{-4} and a weight decay of 10^{-5} . We set the batch size to 512 for vision-only models, and 384 for audio-only and audio-visual models.

Results In Fig. 39a, we plot the recognition mcAP of all models against their total power consumption for the *high-efficiency* tier. We can see that combining vision and audio is better than using only vision or audio. Thus suggests that the two modalities carry complementary cues that are useful for the task. However, all vision-only models outperform their audio-only counterparts, which indicates that vision is the most critical modality for the task. The raw backbones generally perform better than the models using a sampling policy, but at the cost of requiring higher energy, making them impractical to use in online settings. Among the models that use a fixed stride, a lower stride generally improves the performance while hurting energy efficiency. For the vision-only and audio-only backbones, the best stride values within budget are $s = 10$ and $s = 5$, respectively. Among the audio-visual models, AV-LF with a stride of $s = 15$ performs the best among all models that satisfy the budget. Using the greedy or random policy with AV-LF leads to a sharp decline in performance compared to using a fixed stride, showing that sampling very early or randomly in the episode is suboptimal for our online recognition task. AV-cascade also performs worse than most audio-visual models while also requiring more energy, possibly because the audio backbone often outputs wrong but over-confident predictions that prevent switching over to the more reliable vision backbone when required.

For easy reference, in Table 14 we report the recognition performance and total power consumption of our best uni-

modal and audio-visual models within budget for the *high-efficiency* tier.

In Fig. 39b, we plot the recognition mCap of all models against their total power consumption for the *high-performance* tier. Different from the high-efficiency tier, the audio-visual backbone performs worse than the vision-only backbone, possibly because the LaViLa features are strong enough by themselves, and fusing them with audio features through the simple mechanism of linear late fusion reduces their expressivity. Otherwise, the overall behavior of different sampling policies is very similar across the two tiers. We report the recognition performance and total power consumption of the best uni-modal and audio-visual models within the *high-performance* budget in table 15.

Additionally, in Fig. 40, we present a detailed analysis of the keystep labels where the best vision-only model from the *high-efficiency* tier yields the maximum improvement or decline in performance compared to its audio-only counterpart. We observe that the vision-only model produces a large improvement over the audio-only model usually in steps where the activity does not produce distinctive sounds (*e.g.*, *get the fine-mesh sieve, taste the recipe, etc.*). On the other hand, using audio alone helps the most when the activities involve sounds that are strongly indicative of the nature of the task (*e.g.*, *add tomato to a mixture, cut scallions, etc.*).

Finally, we envision that future work on this task will explore more sophisticated *learned* policies, potentially trained using reinforcement learning, in order to adaptively decide when to sample which modality instead of using fixed heuristics. Another promising direction is to investigate efficient transformer-based recognition backbones [175, 193] that can improve recognition performance without significantly affecting the model efficiency.

Contribution statement Tushar Nagarajan led the energy-efficient multimodal benchmark, co-developed the task formulation, and advised the baseline development. Sagnik Majumder led the baseline development effort and contributed all baseline implementations and analysis of results for the benchmark. Mery Ramazanova developed the energy profiler used to evaluate all baselines and contributed to the experimental analysis. Lorenzo Torresani contributed to editing this section. Mitesh Kumar Singh helped design the energy formula. Miao Liu and Shengxin Cindy Zha initiated this benchmark and developed an early version of the task formulation.

13.B.3 Procedure understanding

Annotations For each of the considered procedural tasks, we manually labeled task-graphs as structures encoding the keystep orderings leading to a correct execution of the

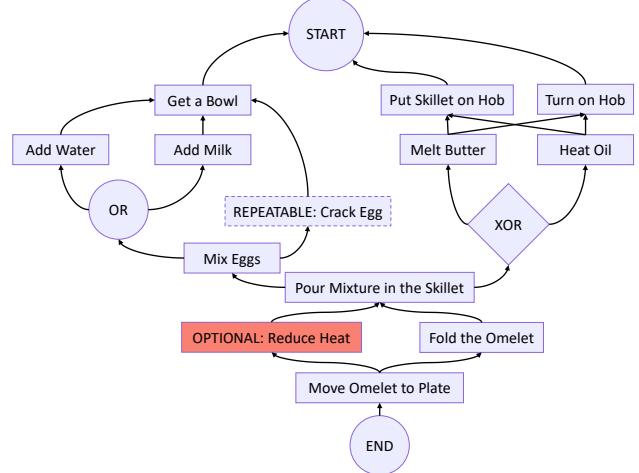


Figure 41. Example task-graph of a "Cooking Omelet" procedure.

procedure.

Task-graphs. We define a task-graph as a directed graph in which nodes represent keysteps and directed edges represent dependencies. For instance, in the task-graph reported in Figure 41, the "Add Milk → Get a Bowl" structure denotes that keystep "Get a Bowl" has to be executed before keystep "Add Milk". Besides directed edges, task-graphs also contain "OR" and "XOR" structures, which combine dependencies logically, as well as "optional" and "repeatable" node attributes.

Task-graph construction. We first familiarized ourselves with the procedural tasks by watching videos with annotated keysteps. We then initialized task graphs with procedural dependencies obtained from keystep annotations through the following procedure: a) a directed graph is first generated from the observed keystep transition frequencies; b) edges of the transition graph are filtered based on transition probabilities using a threshold parameter which is manually tuned for each scenario; c) edge directions are inverted to convert frequent transitions into dependencies. These initial graphs were then refined and corrected by human annotators leveraging common sense and in-domain knowledge.

Segment-level annotations. Let $S = \{s_1, \dots, s_n\}$ be a labeled sequence of keysteps in a given video. We denote with y_i the annotated keystep label of segment s_i and with $Y_{:i} = \{y_1, \dots, y_i\}$ the sequence of labels up to the i -th keystep. Using these keystep annotations, each segment s_i is automatically matched to a task-graph and augmented with the following attributes: 1) a list of *previous keysteps*—these are the in-neighbors of the matched node, 2) *optional labels*—directly derived from the optional node attribute, 3)

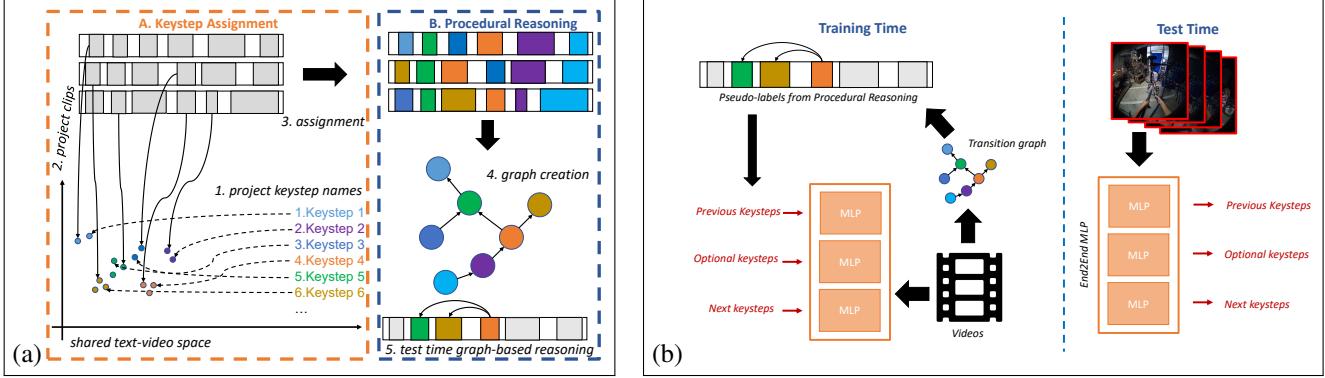


Figure 42. Overview of the two procedure understanding approaches considered in our evaluation: (a) graph-based baselines for procedure understanding rely on a Keystep Assignment and a Procedural Reasoning component; (b) the architecture of our end-to-end baseline.

a *procedural mistake* label—this is set to “true” if the in-neighbors of the matched node do not correspond to segments in the history $Y_{:i}$, 4) the list of *missing keysteps*—the in-neighbors of the matched node not listed in $Y_{:i}$, and 5) the list of *next steps*—nodes for which in-neighbors appear in $Y_{:i}$. Non-repeatable nodes are listed only if they do not appear in $Y_{:i}$.

Formal task definition Given a video segment s_i and its segment history $S_{:i-1} = \{s_1, \dots, s_{i-1}\}$, models have to 1) determine *previous keysteps* (to be performed before s_i); infer if s_i is 2) *optional* or 3) a *procedural mistake*; 4) predict *missing keysteps* (which should have been performed before s_i); and 5) forecast *next keysteps* (for which dependencies are satisfied and hence can be executed next). The task is weakly supervised, with two versions based on the level of supervision: 1) instance-level: segments and their keystep labels are available during training and inference; 2) procedure-level: unlabeled segments and procedure-specific keystep names are given for training and inference.

Note that, when the procedure-level supervision is considered, the input to the model *excludes* keystep labels both at training and test time. At both the procedure and instance levels of supervision, models are required to process the video in a causal fashion, meaning that predictions made at time t only depend on observations made at time $t' < t$.

Baselines We consider two main baseline approaches: graph-based baselines, which leverage an explicit procedure structure mined from videos, and an end-to-end model aimed to make predictions directly from the observed videos.

Graph-based baselines. Graph-based baselines are composed of a keystep assignment and a procedural reasoning component (see Figure 42(a)).

Keystep assignment is applied to obtain a pseudo-labeling of the provided video segments when the supervision is at the procedure-level (i.e., segments are unlabeled and only keystep names are provided). This is achieved by means of an EgoVLPv2 model [82] pre-trained on ego-exo videos and narrations. Video segments and keystep names are projected to the shared video-language space using EgoVLPv2. We hence assigned each video segment to the closest keystep in the representation space according to the cosine distance. In the problem formulation with instance-level supervision (i.e., when keystep labels are available for all segments during both training and testing), we use ground truth labels instead of those obtained from keystep assignment.

The procedural reasoning component creates for each procedure a transition graph based on keystep co-occurrences. In the graph, each node represents a keystep category, while directed edges represent the probability of transitioning from one node to another one. An edge $A \rightarrow B$ is assigned the following weight based on statistics collected from the training videos:

$$P(B|A) = \frac{\text{\# times keystep } B \text{ follows keystep } A}{\text{\# occurrences of keystep } A}$$

At test time, the graph is used to perform procedure understanding and answer the keystep-level questions. Specifically, given current segment s_i : 1) keystep y_{prev} is predicted as the previous keystep with confidence score equal to the transition probability $P(y_i|y_{prev})$, where y_i is the inferred or ground truth keystep label for segment s_i ; 2) segment s_i is predicted as optional based on the empirical probability $\frac{\text{\# training videos containing } y_i}{\text{\# training videos}}$; 3) segment s_i is predicted as a procedural mistake with a score equal to the sum of the transition probabilities to y_i from keysteps y^{prev} that are missing from the keystep history, i.e., $\sum_{y^{prev}} [y^{prev} \notin S_{:i-1}] \cdot P(y_i|y^{prev})$, where $[\cdot]$ is the indicator function; 4) keystep y is predicted as a possible

Supervision	Baseline	Keystep Labels	Prev. Keysteps	Opt. Keysteps	Proc. Mistakes	Miss. Keysteps	Fut. Keysteps
-	Uniform Baseline	-	51.15	50.00	49.17	<u>58.91</u>	60.83
Instance-Level	Graph-Based	Ground Truth	77.24	57.56	63.75	77.53	<u>73.26</u>
Instance-Level	End-to-End	Ground Truth	<u>58.68</u>	<u>60.28</u>	<u>61.31</u>	58.33	73.73
Procedure-Level	Graph-Based	Keystep Assignment	50.80	51.22	55.60	50.58	65.20
Procedure-Level	End-to-End	Keystep Assignment	55.43	66.17	59.28	57.96	71.25

Table 16. Results for the procedure understanding task. Best results are reported in bold, the second best results are underlined. All results are in percentage.

missing keystep with probability $[y_i \notin Y_{:i-1}] \cdot P(y|y_i)$; 5) keystep y is predicted as a future keystep with probability $P(y_i|y)$.

End-to-end baseline. This baseline aims to provide an end-to-end approach to perform procedure understanding directly from the input clip. The baseline predicts previous keysteps, optional keysteps, and next keysteps by feeding video segment features extracted with EgoVLPv2 to three dedicated MLPs. Figure 42(b) illustrates the architecture of the baseline. At training time, MLPs are supervised from the pseudo-labels obtained by graph-based baselines using Mean Squared Error (MSE) score to align the predicted probability distributions to the supervising ones. Missing keysteps and procedural mistakes are predicted from the outputs of the MLP components as in graph-based baselines.

Results We carried out experiments on the following four scenarios: *Install a Wheel*, *Remove a Wheel*, *Clean and Lubricate the Chain*, *First AID - CPR*, *COVID-19 rapid antigen test*. We evaluated our baselines using the calibrated Average Precision (cAP) [27]. Note that, according to this measure, a random baseline would on average achieve a performance of 50%.

Table 16 reports the results obtained by our baselines and compares them against those produced by a “uniform” baseline, predicting previous/optional/mistakes/missing/next keysteps with equal probabilities. Results show that the graph-based baseline relying on ground truth annotations significantly outperforms the uniform baseline for some of the tasks, such as previous and missing keystep predictions. This suggests that even simple keystep co-occurrences are informative to some degree of the overall structure of the procedure. The limited performance gains on optional keystep prediction, procedural mistake detection and future keystep prediction highlight the complexity of the task and the need for further research. The end-to-end model trained with instance-level supervision achieves lower or similar performance, trading accuracy for test-time efficiency, due to the absence of an explicit graph. Procedure-level baselines achieve lower performance due to the limited level of supervision they can rely on. The only exception is for the optional keysteps where the end-to-end baseline outper-

forms the others methods.

Contributions statement Antonino Furnari led the procedure understanding benchmark, and contributed to the task definition, the annotation guidelines, the baseline design, and paper writing. Giovanni Maria Farinella contributed to the task definition, the annotation guidelines, the baseline design, and paper writing. Luigi Seminara contributed to the annotation guidelines, the baseline design, and paper writing; he also developed tools for data annotation, and the baselines for the benchmark. Francesco Ragusa contributed to the annotation guidelines, the baseline design, and paper writing; he also developed tools for data annotation, and the baselines for the benchmark. Kumar Ashutosh contributed to the annotation guidelines, baseline design, and development of data annotation tools. Michael Wray contributed to the task definition, the annotation guidelines, the baseline design, and paper writing. Siddhant Bansal contributed to the task definition, the annotation guidelines, the baseline design, and paper writing. Gene Byrne contributed to the task definition, the annotation guidelines and the baseline design. Tushar Nagarajan contributed to the task definition, and the baseline design. Lorenzo Torresani contributed to editing this section.

13.C. Ego-(Exo) Proficiency Estimation

Annotations The proficiency estimation benchmark consists of two task variants: (1) *demonstrator proficiency estimation*, where the goal is to estimate the absolute skill level of a participant at the task, and (2) *demonstration proficiency estimation*, where the goal is to perform fine-grained analysis of a given task execution to identify good actions from the participant and suggest tips for improvement. We now provide a detailed description of the annotations used in each of these task variants.

Demonstrator proficiency estimation. In this task variant, we assign proficiency labels to each person performing activity demonstrations (one label per person). The proficiency labels span four distinct categories: novice, early expert, intermediate expert, or late expert. Most levels correspond to experts since Ego-Exo4D videos are dominantly targeted towards expert participants who can perform the task successfully (see Sec. 10).

We derive annotations for this task from partici-

	Demonstrator proficiency			Demonstration proficiency		
	Train	Val	Test	Train	Val	Test
Basketball	174	59	87	77	8	29
Bike repair	-	-	-	23	6	9
Cooking	67	9	29	75	10	31
Dance	269	105	138	76	24	28
Health	-	-	-	45	16	4
Music	87	21	26	82	16	24
Rock Climbing	183	67	79	54	29	19
Soccer	9	4	6	2	0	10
Total	789	265	365	434	109	154

Table 17. Distribution over video takes in proficiency estimation benchmark.

part surveys (see Sec. 11.B) and expert commentary (see Sec. 12.A). Participant surveys contain responses to questions about prior experiences in the task such as “How many years have you been doing this task?”, and “Do you have any qualifications/professional training related to the task?” (see Tab. 2 for the complete list). On the other hand, expert commentary is performed by task-specific experts and includes 1 to 10 proficiency scores for each video from the participant (see Sec. 12.A). After consulting with experts hired for each scenario, we designed scenario-specific conversion functions that use the surveys and expert commentaries to produce an estimate of a participant’s proficiency score (see Tab. 18). For example, in basketball and soccer, we use the years of experience to determine skill level since we found this to be an accurate indicator of skill based on analyzing the videos. On the other hand, to determine skill level in bouldering, we use the highest difficulty level of the route solved by the participant. Note that we exclude the bike repair and health scenarios from the demonstrator proficiency task for different reasons. The distribution of participants for bike repair is heavily skewed towards late expert participants, which would heavily bias the task of demonstrator proficiency. The predominant activity in the health collection is COVID testing, where skill levels are hard to determine due to the simplicity of the task. See Fig. 43 for a distribution over proficiency scores within each scenario. We split our dataset into train/val/test splits based on the common split shared across benchmarks. The dataset statistics are shown in Tab. 17.

Demonstration proficiency estimation. For this variant of the task, we leverage temporally localized annotations that include the timestamps of steps demonstrated in the video as well as the proficiency category for each demonstrated step instance (i.e., good execution or a mistake). For this task, we consider all 8 scenarios, as shown in Tab. 17. We derive annotations for this task from expert commentary, where task experts carefully analyze videos and provide timestamped commentary on the participant’s performance

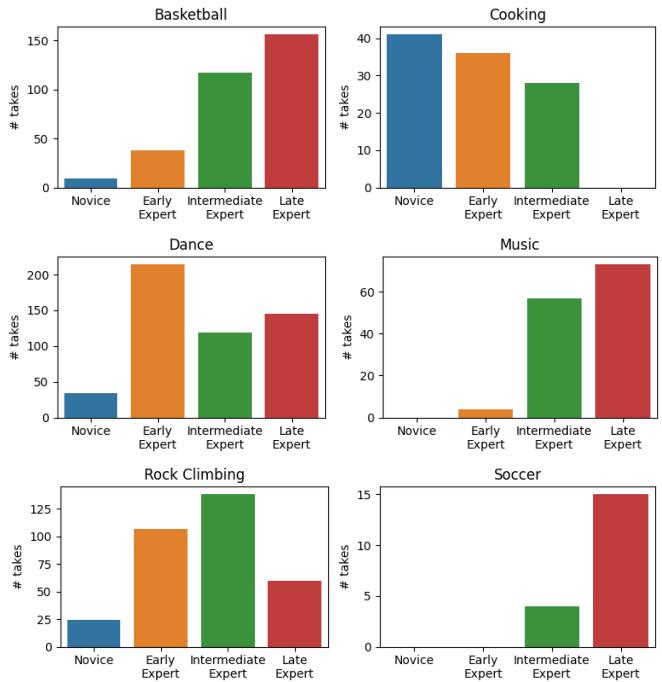


Figure 43. Distribution of demonstrator proficiency scores per scenario.

(see Sec. 12.A). In particular, given a single timestamped comment from an expert, we annotate whether the comment describes a good execution and/or provides tips for improving the participant’s skill level. See Tab. 19 for example annotations. These annotations are then associated with the timestamp provided with each comment to obtain a list of timestamps for good executions $\{t_1^g, t_2^g, \dots\}$ and tips for improvement $\{t_1^i, t_2^i, \dots\}$ in each video. The objective then is to train models that can predict timestamps of good executions or tips for improvement. This amounts to a temporal action localization task for the actions “good executions” and “tips for improvement”. While it would be interesting to build systems to further elaborate what these are (e.g., a specific type of tip, or a natural language explanation), we reserve this for future work. Overall, the demonstration proficiency estimation task consists of 434 train / 109 val / 154 test videos (see Tab. 17 for a breakdown per scenarios).

Formal task definition We now formally define the two problem formulations. In both cases, we are given as input an egocentric video and a set of M exocentric videos of a demonstrator performing a task: $\mathcal{V} = (\mathcal{V}_{ego}, \mathcal{V}_{exo}^{1-M})$. The ego and exo views are time-aligned.

Demonstrator proficiency estimation. The goal here is to estimate the demonstrator’s skill level from one or

Scenario	Novice	Early Expert	Intermediate Expert	Late Expert
Basketball	$X \in [0, 1)$	$X \in [1, 3)$	$X \in [3, 10)$	$X \geq 10$
Soccer	$X \in [0, 1)$	$X \in [1, 3)$	$X \in [3, 10)$	$X \geq 10$
Dancing	$X \in [0, 3)$	$X \in [3, 5)$	$(X \in [5, 10)) \vee ((X \geq 10) \wedge \neg P)$	$(X \geq 10) \wedge P$
Bouldering	$H \leq V3$	$H == V4$	$H == V5$	$H \geq V6$
Music (violin)	$(X \in [0, 3)) \vee (N \in [0, 500))$	$(X \in [3, 5)) \vee (N \in [500, 1000))$	$(X \in [5, 10)) \vee (N \in [1000, 10000))$	$(X \geq 10) \vee (N \geq 10000)$
Music (guitar)	$(X \in [0, 1)) \vee (N \in [0, 500))$	$(X \in [1, 3)) \vee (N \in [500, 1000))$	$(X \in [3, 10)) \vee (N \in [1000, 10000))$	$(X \geq 10) \vee (N \geq 10000)$
Music (piano)	$(X \in [0, 1)) \vee (N \in [0, 500))$	$(X \in [1, 5)) \vee (N \in [500, 1000))$	$(X \in [5, 10)) \vee (N \in [1000, 10000))$	$(X \geq 10) \vee (N \geq 10000)$
Cooking	$P < 3.5$	$P \in [3.5, 5)$	$P \in [5, 8)$	$P \geq 8$

Table 18. **Annotations for demonstrator proficiency estimation.** We designed scenario-specific conversion functions that take in participant surveys and expert commentary assessments to estimate proficiency of participants (i.e., novice, early expert, intermediate expert, and late expert). Legend: X = years of experience performing the task, P = professional training in the task, H = highest difficulty level solved by participant in bouldering, N = estimated number of times performing the task, P = average proficiency rating from expert commentary.

Scenario	Expert comment	Good execution	Tips to improve
Basketball	Nice release. I like the follow through here. You'd like to see the guide hand maybe up a little bit higher on the release of that shot. Maybe to give it better ball control when you're letting go of the shot.	Yes	Yes
Basketball	Great footwork, left foot take off, lifting of the right knee and extending that body up. Love how he's looking up, checking out the backboard, shooting hand behind the basketball. Nice job.	Yes	No
Basketball	He's also really far away from his body and the more he can keep his arm up by his ear, it will give him the most opportunity to make the basket without the defense interrupting.	No	Yes
Bike repair	It's a great method to always double check or do a pre-check before beginning work on a bicycle to make sure the issue that you are working to fix is the only issue that is occurring. If not, you could find a secondary issue or something else that may be greater than the one you are currently working on.	Yes	No
Bike repair	As you can see she clearly slipped on loosening the nut which essentially creates damage to the surface of the nut itself and can round out the nut.	No	Yes
Bouldering	The climber was efficiently able to position herself with one hand on each hold at the start and had, once her hands were positioned, she matched her feet on the hold and efficiently moved to the next hold.	Yes	No
Bouldering	And since she popped out and is swinging out, she can't really keep the tension through her one arm because she's so locked off. So it caused her to kind of just fall off the wall and lose all tension throughout all of her body.	No	Yes
Cooking	You can see there, she's not able to stir properly. She has to push it around, which means that the lime is not gonna be very evenly distributed among the pieces of tomato and cucumber.	No	Yes
Cooking	Using a grinder for fresh pepper is an excellent way to get a lot of flavor. The fresh grind of pepper as opposed to buying already ground pepper really expels the oils and everything in those peppercorns and allows the flavor to be as big as it can possibly be.	Yes	No

Table 19. **Annotations for demonstration proficiency estimation.** We annotated expert comments about a participant's task execution with tags indicating whether each comment describes a good execution or suggests tips for improving skills. Note that the same comment might describe one aspect of the task as being good while suggesting improvements in another aspect (e.g., see row 1).

more task demonstrations. We wish to learn a function f that can estimate the demonstrator's proficiency, i.e., $f(\mathcal{V}) = y^{prof} \in \{\text{novice, early expert, intermediate expert,}$

late expert}\}.

Demonstration proficiency estimation. On the other hand,

the goal in this task variant is to identify parts of the video where the execution was good or needs further improvement. Thus, this time the objective is to learn a function h that can temporally localize instances of good executions and instances of demonstrations that require improvement. Formally, we can express function h as:

$$\hat{G}, \hat{I} = h(\mathcal{V}), \quad (2)$$

where $\hat{G} = \{t_1^g, t_2^g, \dots, t_{|\hat{G}|}^g\}$ are the timestamps where the participant shows good task execution, and $\hat{I} = \{t_1^i, t_2^i, \dots, t_{|\hat{I}|}^i\}$ are the timestamps where the participant needs to improve their skill level.

These tasks inherently benefit from multi-view data. Egocentric video captures fine-grained information about the demonstrator’s hand pose and object interactions, which can be critical to estimating proficiency in tasks that require fine-grained interactions such as cooking (e.g., chopping vegetables) and music (e.g., placement of fingers on the guitar). On the other hand, the exocentric videos provide broader information about the demonstrator’s body pose, which can be highly indicative of proficiency in tasks that require extensive physical motion such as basketball, soccer, and dancing. Our EgoExo4D dataset has 5 views (1 egocentric view, and $M = 4$ exocentric views). We run the proficiency tasks in two settings: one where the exo view is available at test time, and one where it is not. For the latter, benchmarking baseline models with only the egocentric view is important when the target is augmented reality and mobile robotics applications. For the former, benchmarking with both egocentric and exocentric views is informative to capture the multi-view nature of the problem.

Note that the input to the model excludes textual descriptions/narrations of the activity, audio, gaze sensor readings, and any subject information, which would simplify the task significantly at the expense of general applicability since these signals are typically not available for in-the-wild video. We believe the formulation proposed here will encourage the development of video-based methods that (1) do not rely on explicit subject information such as gender, age, ethnicity, etc., and (2) learn to estimate proficiency based on visual cues rather than high-level textual activity descriptions or alternate modalities like sound and eye gaze.

Baselines Next, we define our proposed baselines for each task in the benchmark.

Demonstrator proficiency estimation. We approach this task as a classification problem with four proficiency classes: novice, early expert, intermediate expert, and late expert. We adopt the TimeSformer [13] model for our experiments. TimeSFormer is a video transformer designed for video action recognition/classification that

introduces a novel decoupled spatiotemporal attention mechanism. We train one model on the egocentric view (“ego model”), and a separate model on all 4 exocentric views (“exocentric model”). We resize all videos to 448 pixels along the smallest dimension and use a clip size of 16 frames with a frame rate of 16 FPS. The models are trained to classify individual clips on 8 Quadro RTX 6000 GPUs for 15 epochs. We use the cross-entropy loss as our training objective. At inference time, we employ a late fusion strategy to incorporate information from both egocentric and exocentric video streams. We average the softmax predictions across both egocentric and exocentric models to obtain the final video label prediction. We also average results over three spatial crops during inference following prior work [13]. We report the top-1 accuracy metric.

Demonstration proficiency estimation. We treat this task as a temporal action localization problem with two action categories: ‘good execution’ and ‘needs improvement’. We adopt the ActionFormer [185] model for our experiments. Unlike traditional action localization that defines time windows as outputs, we instead perform timestamp regression since our annotations contain only a single point in time for each good execution or tip for improvement. We accordingly adapt ActionFormer’s post-processing strategy and evaluation metrics. In our task, the predicted timestamps correspond to frames retained after non-maximum suppression (NMS). We remove the regression head of ActionFormer and infer the predicted timestamps from the indices of frames retained after NMS. We also modify the NMS module of ActionFormer to rely on the L_1 -distance between predicted timestamps instead of the tIoU between segments used in [185]. During training, we keep the classification loss from [185] and replace the regression loss with the loss function defined in [72].

We also modify the evaluation setup to use the L_1 -distance between the predicted and ground-truth timestamps instead of tIoU between segments. Therefore, we have metrics thresholded by L_1 -distance in seconds as opposed to tIoU values. We train our models with Omnivore features [42] extracted with a clip size of 32 frames and a stride of 16 frames from all the videos. For the experiments involving multiple views (i.e., multiple exo views or ego + exo views), we simply concatenate the features for all views at each time step.

Results

Demonstrator proficiency estimation. We present results for demonstrator proficiency estimation in Tab. 20. We include two naïve baselines to account for biases in the dataset. The random baseline uniformly samples one skill level at random. The majority-class baseline predicts the majority

Method	Pretrain	Accuracy		
		Ego	Exos	Ego + Exos
Random	-	25.7	25.7	25.7
Majority-class	-	43.0	43.0	43.0
TimeSFormer [13]	-	43.3	40.5	40.8
TimeSFormer [13]	K400	41.9	45.2	45.2
TimeSFormer [13]	EgoVLP	42.7	52.3	51.8
Inference with multiple takes per demonstrator				
TimeSFormer [13]	EgoVLP	46.3	51.0	52.6

Table 20. **Demonstrator proficiency estimation benchmark.** We report top-1 accuracies for various baselines on the demonstrator proficiency estimation task.

class within each scenario. While the random baseline performs poorly, the majority-class baseline achieves higher accuracies, which results from scenario-specific biases in the distribution of expertise levels (see Fig. 43). TimeSFormer trained from random initialization fails to outperform the majority-class baseline. When pre-trained on Kinetics 400 and EgoVLP datasets, TimeSFormer achieves a significant improvement in performance for the exocentric models. However, the ego view alone performs poorly and fails to outperform the naïve majority-class baseline. We found that fusing the predictions from the ego view and exo views does not lead to improved results, likely due to the simplicity of our late-fusion strategy. In the last row of Tab. 20, we further report results when evaluating the model from row 5 on all demonstrations from the same participant (as opposed to a single demonstration). The ego model benefits from analyzing multiple takes from the participant and the fusion of ego and exo views from multiple takes results in the best baseline performance of 52.6%.

We show scenario-specific results in Tab. 21. The model achieves strong performance with only the egocentric view in basketball and cooking, since a close-up view of the objects of interest is essential to assessing skill in these scenarios. On the other hand, the model performs best with the exocentric view as input in dancing since the overall body pose is a primary indicator of proficiency. Furthermore, the TimeSFormer performance in music and bouldering falls well below the majority-class baseline. Overall, our benchmark presents new challenges for video-based skill understanding and our results highlight the difficulty of the task, suggesting good scope for improvement in future work.

Demonstration proficiency estimation. We present results for the demonstration proficiency estimation task in Tab. 22. We include three naïve baselines along with ActionFormer [185]. The “Random tips/good exec.” baseline randomly predicts a tip or a good execution label every 5.97 seconds, i.e., the average temporal span between ad-

Scenario	-	Majority-class		TimeSFormer [13]
		Ego	Exos	Ego + Exos
Basketball	48.27	59.77	52.87	52.87
Cooking	3.44	75.86	65.52	62.06
Dancing	39.13	34.06	52.17	52.17
Music	65.38	42.30	65.38	61.54
Bouldering	48.10	24.05	40.50	40.50
Soccer	83.33	83.33	83.33	83.33

Table 21. **Breakdown of results for demonstrator proficiency estimation across scenarios.** We report top-1 accuracies per scenario for the TimeSFormer model with EgoVLP pre-training.

jacent annotations in our dataset. The “Uniform tips” baseline predicts a tip for improvement label every 5.97 seconds. The “uniform good exec.” baseline predicts a good execution label every 5.97 seconds. We evaluate ActionFormer models trained on ego only, exo only and ego + exo views. All baselines perform poorly on this task and the learned ActionFormer baseline fails to outperform the naïve baselines. This suggests that the task is very challenging and has a significant scope for improvement in methods.

Contribution Statement Santhosh Kumar Ramakrishnan co-led the proficiency estimation benchmark, co-developed the task formulation, and advised the baseline development. Gedas Bertasius co-led the proficiency estimation benchmark, co-developed the task formulation, and advised the baseline development. Arjun Somayazulu developed the demonstrator proficiency estimation baselines. Abrham Gebrselasie developed the demonstration proficiency estimation baselines. Maria Escobar contributed to the task definition and the baseline design. Eugene Byrne contributed to the task definition and the baseline design. Miguel Martin developed an interface for obtaining proficiency estimation scores from the recruited experts. Suyog Jain contributed to an annotation pipeline for demonstration proficiency estimation. Devansh Kukreja built the render flow to generate frame-aligned videos of each camera for each take as model input. Kristen Grauman contributed to the task formulation. Lorenzo Torresani contributed to editing this section.

13.D. Ego Pose

This family of tasks is motivated by recovering the skilled movements of experts in the extreme setting of monocular ego-video input.

13.D.1 3D human body pose from ego-video

Annotations The 3D human body pose annotation process consists of two main stages: (1) automatic ground truth generation, and (2) manual multi-view keypoint annotation/correction. Through this process we derive 3D keypoint annotations for approximately 100K frames.

Method	Ego				Exos				Ego + Exos			
	mAP _{0.25}	mAP _{0.5}	mAP _{1.0}	Avg.	mAP _{0.25}	mAP _{0.5}	mAP _{1.0}	Avg.	mAP _{0.25}	mAP _{0.5}	mAP _{1.0}	Avg.
Random tips/good exec.	0.11	0.51	1.86	0.83	0.11	0.51	1.86	0.83	0.11	0.51	1.86	0.83
Uniform tips	0.19	0.68	2.13	1.00	0.19	0.68	2.13	1.00	0.19	0.68	2.13	1.00
Uniform good exec.	0.12	0.44	1.44	0.66	0.12	0.44	1.44	0.66	0.12	0.44	1.44	0.66
ActionFormer [185]	0.02	0.09	0.17	0.09	0.02	0.12	0.22	0.12	0.02	0.11	0.20	0.11

Table 22. **Demonstration proficiency estimation benchmark.** We report the mean average precision (%) for various baselines on the demonstration proficiency estimation task. mAP_k is measured at an L_1 -distance threshold of k seconds. The average mAP (Avg.) measures the mAP averaged across $k = \{0.25, 0.5, 1.0\}$ seconds.

In the automatic ground truth generation phase, we use off-the-shelf models [23] to predict the 2D bounding boxes from each of the exocentric views. Since there could be multiple people in the scene and we only want to consider the one wearing the egocentric camera, we project the 3D headset location from the MPS output to select which box corresponds to the camera wearer. Then, we run an off-the-shelf 2D human keypoint detector [23] for each bounding box to obtain the 2D keypoints. Finally, we run 3D triangulation with RANSAC to minimize the reprojection errors to obtain the 3D keypoints for the camera wearer.

In the manual annotation phase, we import the undistorted frames and the reprojected 2D keypoints into our multi-view annotation interface.

Formal task definition In our ego-pose task, the goal is to estimate the 3D human pose using either an RGB video input sequence \mathcal{V}_{ego} , an IMU sensor sequence \mathcal{T}_{imu} , or both. For a current frame t , given either an egocentric video $\mathcal{V}_{ego} = \{\mathcal{V}^k, \dots, \mathcal{V}^{t-1}\}$ or an IMU sequence $\mathcal{T}_{imu} = \{\mathcal{T}^k, \dots, \mathcal{T}^{t-1}\}$, where k is a time window in the past, the goal is to predict the human pose P at time t , where $P^t \in \mathbb{R}^{17 \times 3}$ for the 17 annotated joints. Note that at test time, we only estimate the error across the visible annotated joints at frame t . Note that, since the EgoPose benchmark is aimed at promoting the development of methods that perform body pose estimation solely from first-person raw video or IMU data, the input *excludes* egocentric modalities that would unfairly simplify the task (e.g., audio captured from a wearable camera, eye gaze), as well as exocentric video or any signals that can be extracted from it.

Metrics To evaluate the performance of body pose estimation approaches we calculate the Mean Per Joint Position Error (MPJPE) in centimeters (cm), and the Mean Per Joint Velocity Error (MPJVE) in meters per second (m/s).

Baselines We evaluate several baselines methods. Kinpoly [94] is a physics-based simulator that estimates pose using hard physics constraints applied to an underlying artificial humanoid model. EgoEgo [77] is a diffusion based method which maps a sequence of 6 DoF head estimates

to a full sequence of 3D body joint estimates. We also develop a method that combines current IMU-based body pose estimation models Bodiffusion [17] and AvatarPoser [61]. Moreover, to gauge the performance of deep-learning-based methods, we implemented a more straightforward approach. We create a static pose baseline, which consists of fixing the 3D human body pose prediction to be the average pose in the training set and translating it according to the IMU sensor. Thus, the fixed prediction matches the camera location at each frame.

- **Kinpoly.** Kinpoly [94] proposes to use a simulated humanoid to track head pose and create full-body motion based on action types. Based on the input head-pose and action type (out of push, step, avoid, sit, and locomotion), Kinpoly synthesizes realistic human pose and human-object interactions inside a physics simulator. Different from kinematic-based methods that directly output joint angles or positions for pose estimation, Kinpoly outputs joint torques as the final product and controls a simulated humanoid for pose estimation. The laws of physics and human dynamics regulates the output pose to remove artifacts such as jitter and penetration.
- **EgoEgo.** EgoEgo [77] uses a two-step approach for egocentric body pose estimation, by estimating the head pose from the egocentric video first, and then using a diffusion model to generate the full body motion sequence based on the head pose sequence. For head pose estimation, it obtains the initial head pose trajectory using DROID-SLAM [156], and then uses learning-based methods to correct the head pose, including a GravityNet to estimate the additional rotation and a HeadNet with optical flow features as input to estimate the scaling factor to the trajectory. The full body pose is generated with a modified version of DDPM [50] that is conditioned on head pose and trained on AMASS [95]. We show the evaluation of the conditional diffusion part here.
- **IMU-based.** This baseline is inspired by state-of-the-art methods that use transformer-based models for body pose estimation from sparse inputs [17, 61]. We adapt these methods to utilize 3D positions as opposed to the traditional parametric body model. During the training phase, the model was subjected to 40,000 iterations, using the

Method	MPJPE [cm]↓	MPJVE [m/s]↓
static pose	165.78	-
EgoEgo [77]	29.01	1.50
Kinpoly [94]	33.45	1.86
IMU-based [17, 61]	19.60	1.57

Table 23. **Results for the 3D human body pose benchmark.** We report the Mean Per Joint Position Error in cm and the Mean Per Joint Velocity Error in m/s for all the baseline approaches.

Scenario	EgoEgo	Kinpoly	IMU-based
Basketball	28.95	34.90	20.92
Soccer	28.02	32.46	20.68
Bike repair	24.77	22.43	16.01
Health	35.74	-	33.04
Bouldering	40.97	118.88	21.92
Dance	28.71	25.42	19.82
Music	29.99	36.59	13.28

Table 24. **Body pose estimation results per scenario.** The evaluation metric shown is MPJPE in cm.

Adam optimizer with a learning rate of $1e^{-4}$. The window size for temporal analysis was set at 40 frames, and we minimized the Mean Squared Error (MSE) loss between predicted poses and ground truths. As for the input, our model receives a sequence of head poses captured by the device.

Results Table 23 shows the evaluation results of all the baseline approaches. First, note that the static pose baseline obtains a significantly higher MPJPE than all the other approaches. This finding suggests that the poses across different scenarios in the dataset are extremely diverse. Thus, attempting to have the same static pose for all test cases is unfeasible. In contrast, the proposed baseline implementations achieve notable enhancements in performance. Table 24 shows the performance of each method per scenario. Although most scenarios have comparable results, activities like bouldering have a higher error since they require a translation along the vertical axis, which differs from a typical movement in other scenarios. While these developments are promising, we believe that further refinement is possible, especially in lower body pose estimation and to ensure temporal consistency in predictions.

13.D.2 3D human hand pose from ego-video

Annotations As for the case of body pose annotation, the 3D human hand pose annotation process also consists of two stages, i.e., the automatic ground truth generation and the manual multi-view keypoint annotation. Compared to

the body pose, the main difference in automatic ground truth generation is that we also detect hand keypoints from the egocentric frame, and we use the result from the whole body pose estimation to infer the hand locations when there are multiple people in the scene. Similarly, for manual annotation, besides the exocentric frames, we also show the annotators the egocentric frames to allow them to annotate/correct hand keypoints.

For each annotated joint in manual annotations, we provide the number of views used for triangulation as the indicator of the confidence for the provided ground truth data. Meanwhile, the correction the annotators make for hand joints on ego images can serve as the indicator to understand the difficulty for hand reconstruction from the given ego view.

Formal task definition The ego hand pose task entails predicting the three-dimensional coordinates of the camera wearer’s hands from the egocentric frames. Frames from the ego view are extracted and undistorted for both training and evaluation. For baselines requiring 2D hand bounding boxes, we project 3D hand joints in camera coordinates to 2D image planes using the provided intrinsic matrix for the crops.

Note that the inputs of the task do not include depth maps, any additional views from ego or exo cameras, camera pose information, and IMU or active range sensor measurements. We explicitly *exclude* such information for this benchmark to promote the applicability of the methods to general hand pose estimation problems using monocular images.

Metrics To cope with methods estimating wrist-origin and camera-origin hand pose, the ego hand pose baselines are evaluated according to both the MPJPE and the PA-MPJPE metrics. The MPJPE is camera-relative while the PA-MPJPE calculates the average 3D joint distance after performing Procrustes Alignment on wrist-origin hand poses. Both metrics are reported in millimeter (mm) unit.

Baselines We implemented and trained several baseline models for ego hand pose estimation. To estimate the 3D hand joint from monocular 2D ego view images, 2D heatmaps can be explicitly estimated and lifted to 3D space, or 3D joints can be directly estimated from extracted 2D features. The feature extractor backbone could be CNN-based or transformer-based. The proposed baseline methods cover different choices of model designs. All the baseline models work on single frame images without temporal information. The baseline models are trained on manual or manual+automatic annotations, and are only evaluated on manual annotations.

Notably, most baseline methods generate hand mesh as final results in their original paper. We modified them to be trained and supervised only on 2D/3D hand joints (not on hand mesh) to fit the benchmark.

- **THOR-net.** THOR-net [1] uses Keypoint-RCNN as the feature extractor to obtain 2D information and derive 2D hand keypoints heatmaps explicitly. The method then lifts 2D estimates to the 3D space using GraFormer [192], which is a model consisting of Graph Convolutional layers and Attention layers. We use only the 2D-to-3D pose GraFormer branch in THOR-net to adapt the method to our task. The training takes around 4 hours for the manual dataset on a GeForce RTX 4090 Graphics Card, and around 10 hours for the dataset combining manual and automatic annotations.
- **HandOccNet.** HandOccNet [108] uses a ResNet50[48]-based FPN [85] to extract 2D features. The method then uses two Transformer-based modules: Feature Injecting Transformer (FIT) to inject hand information into occluded region, and Self-Enhancing Transformer (SET) to further refine the 2D features. The method proposes a regressor based architecture to produce 2D keypoints, MANO [134] pose, and MANO shape parameters to predict joints and vertices. To accommodate our baseline, only 2D keypoints and 3D joints location losses are used in the training phase. The training takes around 2 hours for the manual dataset on 8 NVIDIA V100 Graphics Cards.
- **POTTER.** POTTER [195] proposes Pooling Attention Transformer (PAT) to extract 2D visual features, which significantly reduces the memory and computational cost without sacrificing performances. The method then applies a mesh regression head HybrIK [76] to generate 3D joint and mesh results. The training takes around 43 minutes for manual dataset, and around 4 hours for manual+auto dataset on a GeForce RTX 4090 Graphics Card.
- **METRO.** METRO [81] extracts a CNN-based global image features. The method then uses a transformer encoder to jointly model vertex-vertex and vertex-joint interactions, and outputs 3D joint coordinates and mesh vertices simultaneously. Since the training of METRO strongly depends on hand mesh supervision, which is not present in the annotations, we borrowed the checkpoint trained on FreiHand [69] dataset and run the inference only, without training it on our benchmark.

Results We report the MPJPE and PA-MPJPE of the baseline models in Table 25, and their corresponding parameter numbers and multiply-accumulate operations (MACs) in Table 26. We further analyze the error distribution across different hand joints. Figure 44 shows that the thumb finger and finger tips tend to have larger errors, most likely because they are occluded or invisible more often.

	Manual		Manual+Auto	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
METRO* [81]	-	24.18	-	24.18
THOR-net [1]	53.69	18.58	50.51	17.21
HandOccNet [108]	52.30	19.15	51.60	19.10
POTTER [195]	40.15	13.31	38.74	13.19

Table 25. MPJPE and PA-MPJPE in mm for ego hand pose baseline models. * denotes methods *not* trained on the benchmark.

	THOR-net[1]	HandOccNet[108]	POTTER[195]
Params (M)	59.5	37.22	14.5
MACs (G)	123.6	15.5	5.2

Table 26. Number of parameters and MACS for the different ego hand pose baselines.

# visible views	2	3	4	5	6
PA-MPJPE (mm)	18.31	14.95	12.95	12.07	10.47

Table 27. PA-MPJPE for joints that are visible in different number of views (including ego and exo views). Results generated from POTTER [195] evaluation.

For each annotated joint, the manual annotations keep record of the number of views where the joint is visible. The visible 2D observation is then used for triangulation in 3D ground truth generation. This can be taken as an indicator of the uncertainty of the ground truth, and the difficulty level for the estimation of the joint (usually, a joint visible by fewer views indicates that it is more entangled with objects or other part of the hand). Table 27 shows that the PA-MPJPE decreases as the visible number of views increases.

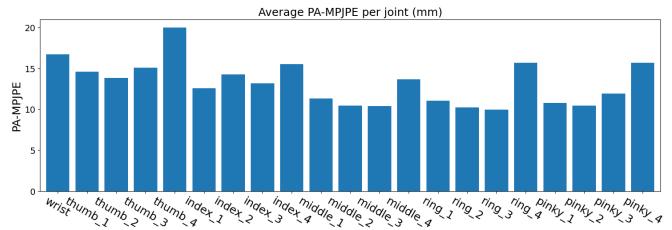


Figure 44. Average PA-MPJPE for each joint. Results generated from POTTER [195] evaluation.

Contribution Statement Kris Kitani co-led the ego-pose benchmark, and provided directional guidance on the task definition, the annotation methodology, and the baseline development. Jianbo Shi co-led the ego-pose benchmark, and provided directional guidance on automatic 3D hand pose generation and development of ego hand pose baseline methods. Maria Escobar led the ego-pose body baseline development, implemented the IMU-based baseline and contributed to experiment analysis. Cristhian Forigua devel-

oped the static pose baseline and contributed to the implementation of the IMU-based baseline. Fu-Jen Chu developed the multi-view annotation UI, the hand pose annotation guidelines, and the data preprocessing code for ego-pose annotation; he also trained and evaluated the HandOc-cNet baseline. Rawal Khirodkar developed the multi-view triangulation and 3D body keypoint estimation pipeline. Zhengyi Luo contributed to the Kinpoly baseline and to the coordinate transform for Aria head poses. Shan Su led the ego-pose hands baseline development, and contributed to automatic 3D hand pose generation, task definition, and annotation development; she also evaluated the baseline using METRO. Suyog Jain developed the annotation pipeline to scale the annotation collection, worked on training the annotators and managed the overall annotation process. Miguel Martin contributed to the automatic ground truth generation pipeline and provided high-level coordination of the body and hands automatic ground truth generation. Jinxu Zhang developed automatic ground truth generation for 3D hand pose; he also trained and evaluated baseline model POTTER. Yiming Huang trained and evaluated the baseline model THOR-net. Zhifan Zhu developed the METRO hand pose baseline method. Jing Huang led the automatic ground truth generation effort, refined body pose annotation guidelines, coordinated ego-pose body baseline development, and ran the EgoEgo body pose baseline. Lorenzo Torresani contributed to editing this section.