Ego4D: Around the World in 3,000 Hours of Egocentric Video


1Meta AI, 2University of Texas at Austin, 3University of Minnesota, 4University of Catania, 5Meta Reality Labs, 6Georgia Tech, 7Carnegie Mellon University, 8UC Berkeley, 9University of Bristol, 10King Abdullah University of Science and Technology, 11National University of Singapore, 12Carnegie Mellon University Africa, 13Meta, 14Universidad de los Andes, 15University of Tokyo, 16Indiana University, 17International Institute of Information Technology, Hyderabad, 18MIT, 19University of Pennsylvania, 20Dartmouth

Abstract

We introduce Ego4D, a massive-scale egocentric video dataset and benchmark suite. It offers 3,670 hours of daily-life activity video spanning hundreds of scenarios (household, outdoor, workplace, leisure, etc.) captured by 931 unique camera wearers from 74 worldwide locations and 19 different countries. The approach to collection is designed to uphold rigorous privacy and ethics standards, with consenting participants and robust de-identification procedures where relevant. Ego4D dramatically expands the volume of diverse egocentric video footage publicly available to the research community. Portions of the video are accompanied by audio, 3D meshes of the environment, eye gaze, stereo, and/or synchronized videos from multiple egocentric cameras at the same event. Furthermore, we present a host of new benchmark challenges centered around understanding the first-person visual experience in the past (querying an episodic memory), present (analyzing hand-object manipulation, audio-visual conversation, and social interactions), and future (forecasting activities). By publicly sharing this massive annotated dataset and benchmark suite, we aim to push the frontier of first-person perception. Project page: https://ego4d-data.org/

1. Introduction

Today’s computer vision systems excel at naming objects and activities in Internet photos or video clips. Their tremendous progress over the last decade has been fueled by major dataset and benchmark efforts, which provide the annotations needed to train and evaluate algorithms on well-defined tasks [49, 60, 61, 92, 108, 143]. While this progress is exciting, current datasets and models represent only a limited definition of visual perception. First, today’s influential Internet datasets capture brief, isolated moments in time from a third-person “spectator” view.
However, in both robotics and augmented reality, the input is a long, fluid video stream from the first-person or “egocentric” point of view—where we see the world through the eyes of an agent actively engaged with its environment. Second, whereas Internet photos are intentionally captured by a human photographer, images from an always-on wearable egocentric camera lack this active curation. Finally, first-person perception requires a persistent 3D understanding of the camera wearer’s physical surroundings, and must interpret objects and actions in a human context—attentive to human-object interactions and high-level social behaviors.

Motivated by these critical contrasts, we present the Ego4D dataset and benchmark suite. Ego4D aims to catalyze the next era of research in first-person visual perception. Ego is for egocentric, and 4D is for 3D spatial plus temporal information.

Our first contribution is the dataset: a massive ego-video collection of unprecedented scale and diversity that captures daily life activity around the world. See Figure 1. It consists of 3,670 hours of video collected by 931 unique participants from 74 worldwide locations in 9 different countries. The vast majority of the footage is unscripted and “in the wild”, representing the natural interactions of the camera wearers as they go about daily activities in the home, workplace, leisure, social settings, and commuting. Based on self-identified characteristics, the camera wearers are of varying backgrounds, occupations, gender, and ages—not solely graduate students! The video’s rich geographic diversity supports the inclusion of objects, activities, and people frequently absent from existing datasets. Since each participant wore a camera for 1 to 10 hours at at time, the dataset offers long-form video content that displays the full arc of a person’s complex interactions with the environment, objects, and other people. In addition to RGB video, portions of the data also provide audio, 3D meshes, gaze, stereo, and/or synchronized multi-camera views that allow seeing one event from multiple perspectives. Our dataset draws inspiration from prior egocentric video data efforts [43,44,129,138,179,201,205,210], but makes significant advances in terms of scale, diversity, and realism.

Equally important to having the right data is to have the right research problems. Our second contribution is a suite of five benchmark tasks spanning the essential components of egocentric perception—indexing past experiences, analyzing present interactions, and anticipating future activity. To enable research on these fronts, we provide millions of rich annotations that resulted from over 250,000 hours of annotator effort and range from temporal, spatial, and semantic labels, to dense textual narrations of activities, natural language queries, and speech transcriptions.

Ego4D is the culmination of an intensive two-year effort by 14 institutions around the world who came together for the common goal of spurring new research in egocentric perception. We are kickstarting that work with a formal benchmark challenge to be held at CVPR 2022. In the coming years, we believe our contribution can catalyze new research not only in vision, but also robotics, augmented reality, 3D sensing, multimodal learning, speech, and language. These directions will stem not only from the benchmark tasks we propose, but also alternative ones that the community will develop leveraging our massive, publicly available dataset.
2. Related Work

Large-scale third-person datasets In the last decade, annotated datasets have both presented new problems in computer vision and ensured their solid evaluation. Existing collections like Kinetics [108], A V A [92], UCF [207], ActivityNet [61], HowTo100M [157], ImageNet [49], and COCO [143] focus on third-person Web data, which have the benefit and bias of a human photographer. In contrast, Ego4D is first-person. Passively captured wearable camera video entails unusual viewpoints, motion blur, and lacks temporal curation. Notably, pre-training egocentric video models with third-person data [70, 221, 224, 239] suffers from the sizeable domain mismatch [139, 201].

Egocentric video understanding Egocentric video offers a host of interesting challenges, such as human-object interactions [26, 46, 163], activity recognition [110, 139, 243], anticipation [4, 75, 86, 144, 205], video summarization [48, 129, 131, 147, 148, 232], detecting hands [16, 134], parsing social interactions [66, 168, 231], and inferring the camera wearer’s body pose [107]. Our dataset can facilitate new work in all these areas and more, and our proposed benchmarks (and annotations thereof) widen the tasks researchers can consider moving forward. We defer discussion of how prior work relates to our benchmark tasks to Sec. 5.

Egocentric video datasets Multiple egocentric datasets have been developed over the last decade. Most relevant to our work are those containing unscripted daily life activity, which includes EPIC-Kitchens [43, 44], UT Ego [129, 210], Activities of Daily Living (ADL) [179], and the Disney dataset [66]. The practice of giving cameras to participants to take out of the lab, first explored in [66, 129, 179], inspires our approach. Others are (semi-)scripted, where camera wearers are instructed to perform a certain activity, as in Charades-Ego [201] and EGTEA [138]. Whereas today’s largest ego datasets focus solely on kitchens [43, 44, 129, 138], Ego4D spans hundreds of environments both indoors and outdoors. Furthermore, while existing datasets rely largely on graduate students as camera wearers [43, 44, 66, 129, 138, 168, 179, 194, 210], Ego4D camera wearers are of a much wider demographic, as detailed below. Aside from daily life activity, prior ego datasets focus on conversation [170], inter-person interactions [66, 168, 194, 231], place localization [183, 208], multimodal sensor data [124, 166, 204], human hands [16, 134] human-object interaction [106, 184], and object tracking [56].

Ego4D is an order of magnitude larger than today’s largest egocentric datasets both in terms of hours of video (3,670 hours vs. 100 in [43]) and unique camera wearers (931 people vs. 71 in [201]); it spans hundreds of environments (rather than one or dozens, as in existing collections); and its video comes from 74 worldwide locations and 9 countries (vs. just one or a few cities). The Ego4D annotations are also of unprecedented scale and depth, with millions of annotations supporting multiple complex tasks. As such, Ego4D represents a step change in dataset scale and diversity. We believe both factors are paramount to pursue the next generation of perception for embodied AI.

3. Ego4D Dataset

Next we overview the dataset, which is publicly available under an Ego4D license.

3.1. Collection strategy and camera wearers

Not only do we wish to amass an ego-video collection that is substantial in scale, but we also want to ensure its diversity of people, places, objects, and activities. Furthermore, for realism, we are interested in unscripted footage captured by people wearing a camera for long periods of time.

To this end, we devised a distributed approach to data collection. The Ego4D project consists of 14 teams from universities and labs in 9 countries and 5 continents (see map in Figure 1). Each team recruited participants to wear a camera for 1 to 10 hours at a time, for a total of 931 unique camera wearers and 3,670 hours of video in this first dataset release (Ego4D-3K). Participants in 74 total cities were recruited by word of mouth, ads, and postings on community bulletin boards. Some teams recruited participants with occupations that have interesting visual contexts, such as bakers, carpenters, landscapers, or mechanics.

Both the geographic spread of our team as well as our approach to recruiting participants were critical to arrive at a diverse demographic composition, as shown in Figure 2.1 Participants cover a wide variety of occupations, span many age brackets, with 96 of them over 50 years old, and 45% are female. Two participants identified as non-binary, and two preferred not to say a gender.

---

1For 64% of all participants; missing demographics are due to protocols or participants opting out of answering specific questions.
Figure 3. Scenarios in Ego4D. Outer circle shows the 14 most common scenarios (70% of the data). Wordle shows scenarios in the remaining 30%. Inner circle is color coded by the contributing partner (see map color legend in Fig 1).

3.2. Scenarios composing the dataset

What activities belong in an egocentric video dataset? Our research is motivated by problems in robotics and augmented reality, where vision systems will encounter daily life scenarios. Hence, we consulted a survey from the U.S. Bureau of Labor Statistics\(^2\) that captures how people spend the bulk of their time in the home (e.g., cleaning, cooking, yardwork), leisure (e.g., crafting, games, attending a party), transportation (e.g., biking, car), errands (e.g., shopping, walking dog, getting car fixed), and in the workplace (e.g., talking with colleagues, making coffee).

To maximize coverage of such scenarios, our approach is a compromise between directing camera wearers and giving no guidance at all: (1) we recruited participants whose collective daily life activity would naturally encompass a spread of the scenarios (as selected freely by the participant), and (2) we asked participants to wear the camera at length (at least as long as the battery life of the device) so that the activity would unfold naturally in a longer context. A typical raw video clip in our dataset lasts 8 minutes—significantly longer than the 10 second clips often studied in third-person video understanding [108]. In this way, we capture unscripted activity while being mindful of the scenarios’ coverage.

The exception is for certain multi-person scenarios, where we asked participants at five sites who had consented to share their conversation audio and unblurred faces to take part in social activities, such as playing games. We leverage this portion of Ego4D for the Audio-Visual and Social Interaction benchmarks (Sec. 5.3 and 5.4).

Figure 3 shows the wide distribution of scenarios captured in our dataset. Note that within each given scenario there are typically dozens of actions taking place, e.g., the carpentry scenario includes hammering, drilling, moving wood, etc. Overall, the 931 camera wearers bestow our dataset with a glimpse of daily life activity around the world.

\(^2\)https://www.bls.gov/news.release/atus.nr0.htm

3.3. Cameras and modalities

To avoid models overfitting to a single capture device, seven different head-mounted cameras were deployed across the dataset: GoPro, Vuzix Blade, Pupil Labs, ZShades, OR-DRO EP6, iVue Rincon 1080, and Weeview. They offer tradeoffs in the modalities available (RGB, stereo, gaze), field of view, and battery life. The field of view and camera mounting are particularly influential: while a GoPro mounted on the head pointing down offers a high resolution view of the hands manipulating objects (Fig. 5, right), a heads-up camera like the Vuzix shares the vantage of a person’s eyes, but will miss interactions close to the body (Fig. 5, left).

In addition to video, portions of Ego4D offer several other data modalities: 3D scans, audio, gaze\(^3\), stereo, multiple synchronized wearable cameras, and textual narrations. See Table 1. Each can support new research challenges. For example, having Matterport3D scans of the environment coupled with ego-video clips (Figure 4) offers a unique opportunity for understanding dynamic activities in a persistent 3D context, as we exploit in the Episodic Memory benchmark (see Sec. 5.1). Multiple synchronized egocentric video streams allow accounting for the first and second-person view in social interactions. Audio allows analysis of conversation and acoustic scenes and events.

3.4. Privacy and ethics

From the onset, privacy and ethics standards were critical to this data collection effort. Each partner was responsible for developing a policy. While specifics vary per site, this generally entails:

- Comply with own institutional research policy, e.g., independent ethics committee review where relevant

---

\(^3\)Eye trackers were deployed by Indiana U. and Georgia Tech only.
Table 1. Modalities of data in Ego4D and their amounts. “Narrations” are dense, timestamped descriptions of camera wearer activity (cf. Sec. 4). “3D scans” are meshes from Matterport3D scanners for the full environment in which the video was captured. “Faces” refers to video where participants consented to remain unblurred. “Multi-cam” refers to synchronized video captured at the same event by multiple camera wearers. “Features” refers to precomputed SlowFast [70] video features.

<table>
<thead>
<tr>
<th>Modality</th>
<th>RGB video</th>
<th>Text narrations</th>
<th>Features</th>
<th>Audio</th>
<th>Faces</th>
<th>3D scans</th>
<th>Stereo</th>
<th>Gaze</th>
<th>IMU</th>
<th>Multi-cam</th>
</tr>
</thead>
<tbody>
<tr>
<td># hours</td>
<td>3,670</td>
<td>3,670</td>
<td>3,670</td>
<td>2,535</td>
<td>612</td>
<td>491</td>
<td>80</td>
<td>45</td>
<td>836</td>
<td>224</td>
</tr>
</tbody>
</table>

Obtain informed consent of camera wearers, who can ask questions and withdraw at any time, and are free to review and redact their own video
• Respect rights of others in private spaces, and avoid capture of sensitive areas or activities
• Follow de-identification requirements for personally identifiable information (PII)

In short, these standards typically require that the video be captured in a controlled environment with informed consent by all participants, or else in public spaces where faces and other PII are blurred. Appendix K in the supplementary materials discusses potential negative societal impact.

3.5. Possible sources of bias

While Ego4D pushes the envelope on massive everyday video from geographically and demographically diverse sources, we are aware of a few biases in our dataset. 74 locations is still a long way from complete coverage of the globe. In addition, the camera wearers are generally located in urban or college town areas. The COVID-19 pandemic led to ample footage in stay-at-home scenarios such as cooking, cleaning, crafts, etc. and more limited opportunities to collect video at major social public events. In addition, since battery life prohibits daylong filming, the videos—though unscripted—tend to contain more active portions of a participant’s day. Finally, Ego4D annotations are done by crowdsourced workers in two sites in Africa. This means that there will be at least subtle ways in which the language-based narrations are biased towards their local word choices.

4. Narrations of Camera Wearer Activity

Before any other annotation occurs, we pass all video through a narration procedure. Inspired by the pause-and-talk narrator [44], annotators are asked to watch a 5 minute clip of video, summarize it with a few sentences, and then re-watch, pausing repeatedly to write a sentence about each thing the camera wearer does. We record the timestamps and the associated free-form sentences. See Figure 5. Each video receives two independent narrations from different annotators. The narrations are temporally dense: on average we received 13.2 sentences per minute of video, for a total of 3.85M sentences. In total the narrations describe the Ego4D video using 1,772 unique verbs (activities) and 4,336 unique nouns (objects). See Appendix D for details.

The narrations allow us to (1) perform text mining for data-driven taxonomy construction for actions and objects, (2) sort the videos by their content to map them to relevant benchmarks, and (3) identify temporal windows where certain annotations should be seeded. Beyond these uses, the narrations are themselves a contribution of the dataset, potentially valuable for research on video with weakly aligned natural language. To our knowledge, ours is the largest repository of aligned language and video (e.g., HowTo100M [157], an existing Internet repository with narrations, contains noisy spoken narrations that only sometimes comment on the activities taking place).

5. Ego4D Benchmark Suite

First-person vision has the potential to transform many applications in augmented reality and robotics. However, compared to mainstream video understanding, egocentric perception requires new fundamental research to account for long-form video, attention cues, person-object interactions, multi-sensory data, and the lack of manual temporal curation inherent to a passively worn camera.

Inspired by all these factors, we propose a suite of challenging benchmark tasks. The five benchmarks tackle the past, present, and future of first-person video. See Figure 6. The following sections introduce each task and its annotations. The first dataset release has annotations for 48-1,000 hours of data per benchmark, on top of the 3,670 hours of data that is narrated. The Appendices describe how we sampled videos per benchmark to maximize relevance to the task while maintaining geographic diversity.

We developed baseline models drawing on state-of-the-art components from the literature in order to test drive all Ego4D benchmarks. The Appendices present the baseline models and quantitative results. We are running a formal Ego4D competition at CVPR 2022 inviting the research community to improve on these baselines.
5.1. Episodic Memory

**Motivation**  Egocentric video from a wearable camera records the who/what/when/where of an individual’s daily life experience. This makes it ideal for what Tulving called episodic memory [213]: specific first-person experiences (“what did I eat and who did I sit by on my first flight to France?”), to be distinguished from semantic memory (“what’s the capital of France?”). An augmented reality assistant that processes the egocentric video stream could give us super-human memory if it could appropriately index our visual experience and answer queries.

**Task definition**  Given an egocentric video and a query, the Ego4D Episodic Memory task requires localizing where the answer can be seen within the user’s past video. We consider three query types. (1) **Natural language queries** (NLQ), in which the query is expressed in text (e.g., “What did I put in the drawer?”), and the output response is the temporal window where the answer is visible or deducible. (2) **Visual queries** (VQ), in which the query is a static image of an object, and the output response localizes the object the last time it was seen in the video, both temporally and spatially. The spatial response is a 2D bounding box on the object, and optionally a 3D displacement vector from the current camera position to the object’s 3D bounding box. VQ captures how a user might teach the system an object with an image example, then later ask for its location (“Where is this [picture of my keys]?”). (3) **Moments queries** (MQ), in which the query is the name of a high-level activity or “moment”, and the response consists of all temporal windows where the activity occurs (e.g., “When did I read to my children?”). See Figure 7.

**Annotations**  For language queries, we devised a set of 13 template questions meant to span things a user might ask to augment their memory, such as “what is the state of object X?”, e.g., “did I leave the window open?”. Annotators express the queries in free-form natural language, and also provide the slot filling (e.g., X = window). For moments, we established a taxonomy of 110 activities in a data-driven, semi-automatic manner by mining the narration summaries. Moments capture high-level activities in the camera wearer’s day, e.g., *setting the table* is a moment, whereas *pick up* is an action in our Forecasting benchmark (Sec. 5.5).

For NLQ and VQ, we ask annotators to generate language/visual queries and couple them with the “response track” in the video. For MQ, we provide the taxonomy of labels and ask annotators to label clips with each and every temporal segment containing a moment instance. In total, we have ~74K total queries spanning 1,000 hours of video.

**Evaluation metrics and baselines**  For NLQ, we use top-k recall at a certain temporal intersection over union (tIoU) threshold. MQ adopts a popular metric used in temporal action detection: mAP at multiple tIoU thresholds, as well as top-kx recall. VQ adopts temporal and spatio-temporal localization metrics as well as timeliness metrics that encourage speedy searches. Appendix F presents the baseline models we developed and reports results.

**Relation to existing tasks**  Episodic Memory has some foundations in existing vision problems, but also adds new challenges. All three queries call for spatial reasoning in a static environment coupled with dynamic video of a person who moves and changes things; current work largely treats these two elements separately. The timeliness metrics encourage work on intelligent contextual search. While current literature on language+vision focuses on captioning and question answering for isolated instances of Internet data [12, 35, 119, 228], NLQ is motivated by queries about the camera wearer’s own visual experience and operates over long-term observations. VQ upgrades object instance recog-
whether an object state change has taken place or not.

wearer changes the state of an object by using or manipulating it—which we call an object state change. Though cutting a piece of lumber in half can be achieved through many methods (e.g., various tools, force, speed, grasps, end-effectors), all should be recognized as the same state change. This generalization ability will enable us to understand human actions better, as well as to train robots to learn from human demonstrations in video.

5.2. Hands and Objects

Motivation While Episodic Memory aims to make past video queryable, our next benchmark aims to understand the camera wearer’s present activity—in terms of interactions with objects and other people. Specifically, the Hands and Objects benchmark captures how the camera wearer changes the state of an object by using or manipulating it—which we call an object state change. Though cutting a piece of lumber in half can be achieved through many methods (e.g., various tools, force, speed, grasps, end-effectors), all should be recognized as the same state change. This generalization ability will enable us to understand human actions better, as well as to train robots to learn from human demonstrations in video.

Task definitions We interpret an object state change to include various physical changes, including changes in size, shape, composition, and texture. Object state changes can be viewed along temporal, spatial and semantic axes, leading to these three tasks: (1) Point-of-no-return temporal localization: given a short video clip of a state change, the goal is to estimate the keyframe that contains the point-of-no-return (PNR) (the time at which a state change begins); (2) State change object detection: given three temporal frames (pre, post, PNR), the goal is to regress the bounding box of the object undergoing a state change; (3) Object state change classification: given a short video clip, the object is to classify whether an object state change has taken place or not.

Annotations We select the data to annotate based on activities that are likely to involve hand-object interactions (e.g., knitting, carpentry, baking, etc.). We start by labeling each narrated hand-object interaction. For each, we label three moments in time (pre, PNR, post) and the bounding boxes for the hands, tools, and objects in each of the three frames. We also annotate the state change types (remove, burn, etc., see Fig. 8), action verbs, and nouns for the objects.

Evaluation metrics and baselines Object state change temporal localization is evaluated using absolute temporal error measured in seconds. Object state change classification is evaluated by classification accuracy. State change object detection is evaluated by average precision (AP). Appendix G details the annotations and presents baseline model results for the three Hands and Objects tasks.

Relation to existing tasks Limited prior work considers object state change in photos [102,164] or video [8,68,242]; Ego4D is the first video benchmark dedicated to the task of understanding object state changes. The task is similar to action recognition (e.g., [100,110,139,221,243]) because in some cases a specific action can correspond to a specific state change. However, a single state change (e.g., cutting) can also be observed in many forms (various object-tool-action combinations). It is our hope that the proposed benchmarks will lead to the development of more explicit models of object state change, while avoiding approaches that simply overfit to action or object observations.

5.3. Audio-Visual Diarization

Motivation Our next two tasks aim to understand the camera wearer’s present interactions with people. People communicate using spoken language, making the capture of conversational content in business meetings and social settings a problem of great scientific and practical interest. While diarization has been a standard problem in the speech recognition community, Ego4D brings in two new aspects (1) simultaneous capture of video and audio (2) the egocentric perspective of a participant in the conversation.

Task definition and annotations The Audio-Visual Diarization (AVD) benchmark is composed of four tasks (see Figure 9):

- Localization and tracking of the participants (i.e., candidate speakers) in the visual field of view (FoV). A bounding box is annotated around each participant’s face.
- Active speaker detection where each tracked speaker is assigned an anonymous label, including the camera wearer who never appears in the visual FoV.
- Diarization of each speaker’s speech activity, where we provide the time segments corresponding to each speaker’s voice activity in the clip.
- Transcription of each speaker’s speech content (only English speakers are considered for this version).

Evaluation metrics and baselines We use standardized object tracking (MOT) metrics [18,19] to evaluate speaker localization and tracking in the visual FoV. Speaker detection with anonymous labels is evaluated using the speaker
5.4. Social Interactions

Motivation An egocentric video provides a unique lens for studying social interactions because it captures utterances and nonverbal cues [115] from each participant’s unique view and enables embodied approaches to social understanding. Progress in egocentric social understanding could lead to more capable virtual assistants and social robots. Computational models of social interactions can also provide new tools for diagnosing and treating disorders of socialization and communication such as autism [188], and could support novel prosthetic technologies for the hearing-impaired.

Task definition While the Ego4D dataset can support such a long-term research agenda, our initial Social benchmark focuses on multimodal understanding of conversational interactions via attention and speech. Specifically, we focus on identifying communicative acts that are directed towards the camera-wearer, as distinguished from those directed to other social partners: (1) Looking at me (LAM): given a video in which the faces of social partners have been localized and identified, classify whether each visible face is looking at the camera wearer; and (2) Talking to me (TTM): given a video and audio segment with the same tracked faces, classify whether each visible face is talking to the camera wearer.

Annotations Social annotations build on those from AV diarization (See 5.3). Given (1) face bounding boxes labeled with participant IDs and tracked across frames, and (2) associated active speaker annotations that identify in each frame whether the social partners whose faces are visible are speaking, annotators provide the ground truth labels for LAM and TTM as a binary label for each face in each frame. For LAM, annotators label the time segment (start and end time) of a visible person when the individual is looking at the camera wearer. For TTM, we use the vocal activity annotation from AVD, then identify the time segment when the speech is directed at the camera wearer. See Figure 9.

Evaluation metrics and baselines We use mean average precision (mAP) and Top-1 accuracy to quantify the classification performance for both tasks. Unlike AVD, we measure precision at every frame. Appendix I provides details and presents Social baseline models and results.

Relation to existing tasks Compared to [67], Ego4D contains substantially more participants, hours of recording, and variety of sensors and social contexts. The LAM task is most closely related to prior work on eye contact detection in egovideo [36, 159], but addresses more diverse and challenging scenarios. Mutual gaze estimation [54, 150–152, 172, 176] and gaze following [37, 65, 111, 186] are also relevant. The TTM task is related to audio-visual speaker detection [7, 193] and meeting understanding [21, 132, 154].

5.5. Forecasting

Motivation Having addressed the past and present of the camera wearer’s visual experience, our last benchmark moves on to anticipating the future. Forecasting movements and interactions requires comprehending the camera wearer’s intention. It has immediate applications in AR and human-robot interaction, such as anticipatively turning on appliances or moving objects for the human’s convenience. The scientific motivation can be seen by analogy with language models such as GPT-3 [24], which implicitly capture knowledge needed by many other tasks. Rather than predict the next word, visual forecasting models the dynamics of an agent acting in the physical world.

Task definition The Forecasting benchmark includes four tasks (Fig. 10): (1) Locomotion prediction: predict a set of possible future ground plane trajectories of the camera wearer. (2) Hand movement prediction: predict the hand
Figure 10. The Forecasting benchmark aims to predict future locomotion, movement of hands, next object interactions, and sequences of future actions.

positions of the camera wearer in future frames. (3) Short-term object interaction anticipation: detect a set of possible future interacted objects in the most recent frame of the clip. To each object, assign a verb indicating the possible future interaction and a “time to contact” estimate of when the interaction is going to begin. (4) Long-term action anticipation: predict the camera wearer’s future sequence of actions.

Annotations Using the narrations, we identify the occurrence of each object interaction, assigning a verb and a target object class. The verb and noun taxonomies are seeded from the narrations and then hand-refined. For each action, we identify a contact frame and a pre-condition frame in which we annotate bounding boxes around active objects. The same objects as well as hands are annotated in three frames preceding the pre-condition frame by 0.5s, 1s and 1.5s. We obtain ground truth ego-trajectories of the camera wearer using structure from motion.

Evaluation metrics and baselines We evaluate future locomotion movement and hand movement prediction using L2 distance. Short-term object interaction anticipation is evaluated using a Top-5 mean Average Precision metric which discounts the Top-4 false negative predictions. Long-term action anticipation is evaluated using edit distance. Appendix J details the tasks, annotations, baseline models, and results.

Relation to existing tasks Predicting future events has increasing interest [191]. Previous work considers future localization [113, 120, 174, 230], action anticipation [76, 77, 86, 118, 127, 219], next active object prediction [20, 74], future event prediction [149, 167], and future frame prediction [145, 146, 153, 215, 218, 227]. Whereas past work relies on different benchmarks and task definitions, we propose a unified benchmark to assess progress in the field.

6. Conclusion

Ego4D is a first-of-its-kind dataset and benchmark suite aimed at advancing multimodal perception of egocentric video. Compared to existing work, our dataset is orders of magnitude larger in scale and diversity. The data will allow AI to learn from daily life experiences around the world—seeing what we see and hearing what we hear—while our benchmark suite provides solid footing for innovations in video understanding that are critical for augmented reality, robotics, and many other domains. We look forward to the research that will build on Ego4D in the years ahead.

Contribution statement

Project led and initiated by Kristen Grauman. Program management and operations led by Andrew Westbury. Scientific advising by Jitendra Malik. Authors with stars (*) were key drivers of implementation, collection, and/or annotation development throughout the project. Authors with daggers (†) are faculty PIs and working group leads in the project. The benchmarks brought together many researchers from all institutions including cross-institution baseline evaluations. The Appendices detail the contributions of individual authors for the various benchmarks.

Appendix A provides details about the data collection done per site and acknowledges the primary contributors. The video collected by Meta Reality Labs used Vuzix Blade® Smart Glasses and was done in a closed environment in Meta’s buildings by paid participants who signed consents to share their data. All other video collection and participant recruitment was managed by the university partners. The annotation effort was led by Meta AI.

Acknowledgements We gratefully acknowledge the following colleagues for valuable discussions and support of our project: Aaron Adcock, Andrew Allen, Behrouz Behmardi, Serge Belongie, Antoine Bordes, Mark Broyles, Xiao Chu, Samuel Clapp, Irene D’Ambra, Peter Dodds, Jacob Donley, Ruohan Gao, Tal Hassner, Ethan Henderson, Jiabo Hu, Guillaume Jeanneret, Sanjana Krishnan, Devansh Kukreja, Tsung-Yi Lin, Bobby Otillar, Manohar Paluri, Maja Pantic, Lucas Pinto, Vivek Roy, Jerome Pesenti, Joelle Pineau, Luca Sbordone, Rajan Subramanian, Helen Sun, Mary Williamson, and Bill Wu. We also acknowledge Jacob Chalk for setting up the Ego4D AWS backend and Prasanna Sridhar for developing the Ego4D website. Thank you to the Common Visual Data Foundation (CVDF) for hosting the Ego4D dataset. The universities acknowledge the usage of commercial software for de-identification of video. brighter.ai was used for redacting videos by some universities. Personal data from the U. Bristol was protected by Primloc’s Secure Redact software.

UNICT is supported by MIUR AIM - Attrazione e MobilitaInternazionale Linea 1 - AIM1893589 - CUP E6411800254000. Bristol is supported by UKRI Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Program (DTP), EPSRC Fellowship UMPIRE (EP/TP04391/1). KAUST is supported by the KAUST Office of Sponsored Research through the Visual Computing Center (VCC) funding. National University of Singapore is supported by Mike Shou’s Start-Up Grant. Georgia Tech is supported in part by NSF 2033413 and NIH R01MH114999.
References

[1] Github repository of the ESPNet model zoo. https://github.com/espnet/espnet_model_zoo. We used the ShinjilWatanabe/gigaspeech_asr_train_asr_raw_en_bpe5000_valid.acc.ave model. 53


[28] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec,


Ivan Rodin, Antonino Furnari, Dimitrios Mavroeditis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 2021. 9


A. Senocak, T.-H. Oh, J. Kim, M. Yang, and I. S. Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *TPAMI*, 2019. 8, 45

Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In CVPR, 2020. 40, 41


[228] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 6


