Ego4D: Around the World in 3,600 Hours of Egocentric Video

Kristen Grauman, *Fellow, IEEE*, Andrew Westbury, Eugene Byrne, Vincent Cartillier, Zachary Chavis, Antonino Furnari, *Senior Member, IEEE*, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Devansh Kukreja, 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, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie 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 Puentes, 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, *Fellow, IEEE*, 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, *Fellow, IEEE*

***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 9 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.**

***Index Terms*—Video understanding, egocentric video, first-person vision, datasets and benchmarks.**

I. 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 [32], [39], [40], [58], [65], [85].

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 "spectactor" 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, framing them intentionally to convey a message or capture a memory 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 Fig. 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

Manuscript received 27 January 2023; revised 26 May 2023; accepted 6 August 2023. Date of publication 26 July 2024; date of current version 3 October 2025. Recommended for acceptance by K. Dana, G. Hua, S. Roth, D. Samaras, and R. Singh. *(Corresponding author: Kristen Grauman.)*

Please see the Acknowledgment section of this article for the author affiliations.

Project page: https://ego4d-data.org/

Digital Object Identifier 10.1109/TPAMI.2024.3381075

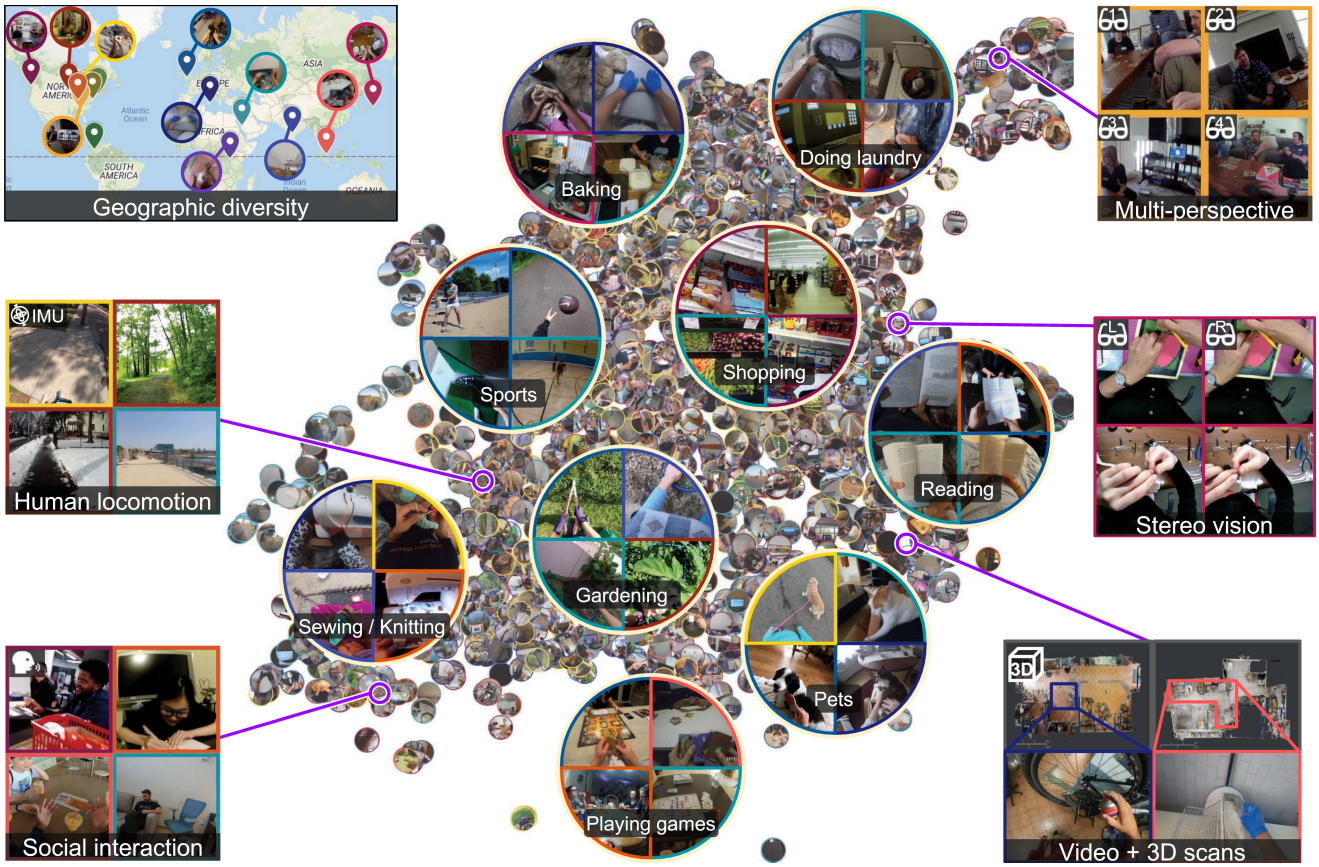


Fig. 1. Ego4D is a massive-scale egocentric video dataset of daily life activity spanning 74 locations worldwide. Here we see a snapshot of the dataset (5% of the clips, randomly sampled) highlighting its diversity in geographic location, activities, and modalities. The data includes social videos where participants consented to remain unblurred. See https://ego4d-data.org/fig1.html for interactive figure.

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 [27], [28], [74], [80], [113], [129], [132], [134], 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.

In particular, we present five tasks: *episodic memory*, in which a long video history becomes queryable for objects and natural language questions; *forecasting*, in which the future actions and trajectories of the camera wearer are anticipated; *hands and objects*, in which the state change for an object being manipulated by the camera wearer is recognized; *audio-visual diarization*, in which the voice activity of who said what, and when, is automatically extracted; and *social interaction*, in which attentional signals of whom is speaking to or looking at whom are inferred. 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 three-year effort by 14 institutions around the world who came together for the common goal of spurring new research in egocentric perception. To kick start work in that direction, so far our team has hosted four benchmark challenges at premier computer vision conferences (CVPR 2022, ECCV 2022, CVPR 2023, and CVPR 2024), where industrial and academic teams around the world competed on 16 tasks from the benchmarks described above. The results from the community are promising: performance has increased beyond our original baselines by as much as 300%! A total of 178 teams from around the world have thus far competed in our formal challenges, and there was a 600% increase in participation between our first offering and the most recent one.

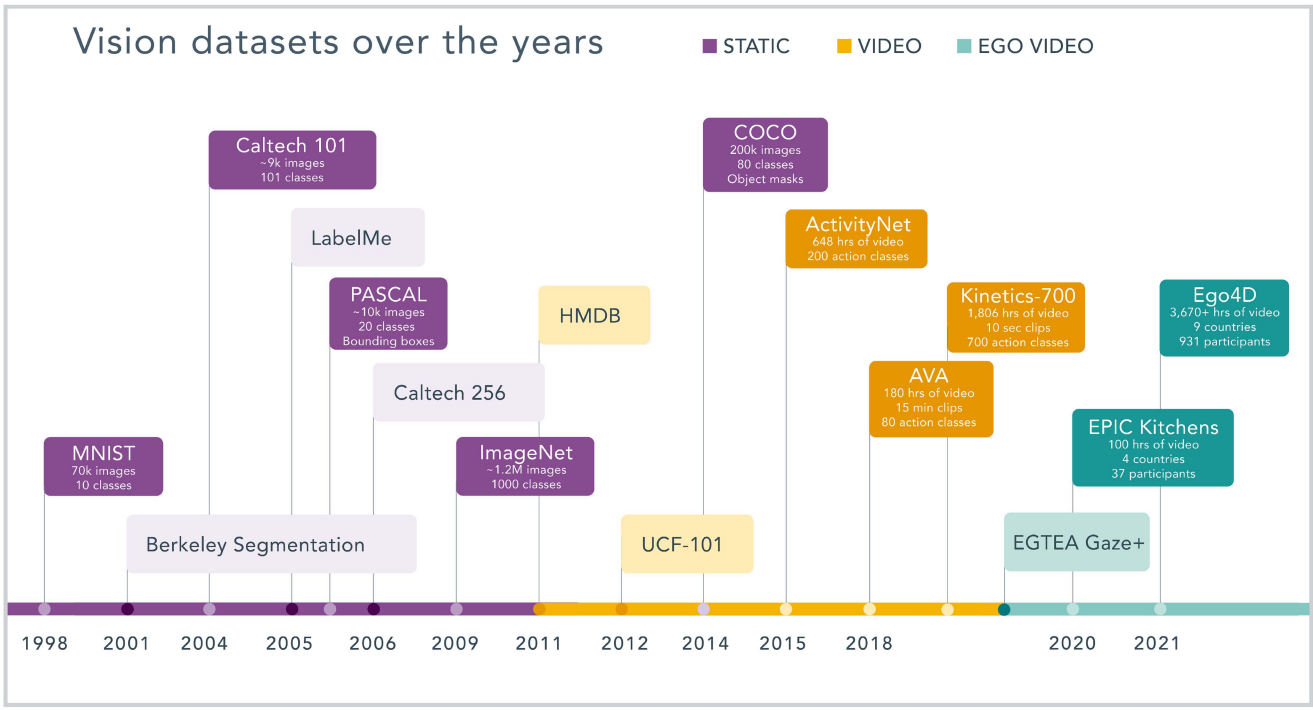


Fig. 2. Over the last two decades, benchmarks and datasets have been instrumental for catalyzing progress in computer vision, providing researchers with a testbed of real-world images and videos as well as rigorous, reproducible evaluation protocols. Ego4D adds to this rich trajectory with a video collection of unprecedented scale and diversity, together with standardized tasks for embodied AI in video.

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. See Fig. 2.

II. 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 [65], AVA [58], ActivityNet [40], HowTo100M [95], ImageNet [32], and COCO [85] 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. The status quo is to pre-train video models with third-person video [46], [138]; however, pre-training egocentric action recognition models with third-person data suffers a domain mismatch and produces significantly worse results than pre-training with first-person data [81], [129].

Egocentric Video Understanding: Egocentric video offers a host of interesting challenges, such as human-object interactions [15], [30], [100], [140], activity recognition [66], [81], [141], [155], anticipation [55], [132], video summarization [31], [75], detecting hands [8], [77], parsing social interactions [43], [106], [146], and inferring the camera wearer's body pose [64]. 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 Section VI.

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 [27], [28], UT Ego [74], [134], Activities of Daily Living (ADL) [113], and the Disney dataset [43]. The practice of giving cameras to participants to take out of the lab, first explored in [43], [74], [113], inspires our approach. Others are (semi-)scripted, where camera wearers are instructed to perform a certain activity, as in Charades-Ego [129] and EGTEA [80]. Whereas today's largest ego datasets focus solely on kitchens [28], [28], [72], [80], Ego4D spans hundreds of environments both indoors and outdoors. Furthermore, while existing datasets rely largely on graduate students as camera wearers [27], [28], [43], [74], [74], [80], [106], [113], [125], [134], Ego4D camera wearers are of a much wider demographic, as detailed below. An exception is Charades-Ego, which recruits Mechanical Turkers to record video in their homes. Aside from daily life activity, prior ego datasets focus on conversation [108], inter-person interactions [43], [106], [125], [146], place localization [117], [133], multimodal sensor data [72], [104], [131], human hands [8], [77] human-object interaction [63], [118], and object tracking [37].

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 [27]) and unique camera wearers (931 people vs. 71 in [129]); 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

TABLE I
COMPARISON BETWEEN EGO4D AND OTHER FIRST-PERSON DATASETS

												Hours					
Datasets	**Year**	**Unscr.**	**Nat. Env.**	**Subj.**	**Loc.**	**Cntr.**	**Scen.**	**Tsk.**	**Dev.**	**3D Env**	**Vids**	**Avg. Len (m)**	**Total**	**Audio**	**Gaze**	**Stereo**	**IMU**	**Sync.**
Ego4D v1	2021	Yes	Yes	931	74	9	136	16	14	15	9,645	24.11	3670	2533.16	33.35	79.82	866.22	497.99
Ego4D v2	2022	Yes	Yes	931	74	9	136	16	14	15	9,655	24.11	3,891	2548.80	33.35	79.82	866.22	497.99
Assembly101 [126]	2022	No	No	53	1	1	1	4	1	0	1425	7.1	167	0	0	0	0	0
MECCANO [118]	2021	No	No	20	2	2	1	5	1	0	20	20.79	6.83	0	6.83	0	0	0
EgoBody [149]	2021	No	No	36	15	1	1	1	1	9	125	2.66	1.84	0	1.84	0	0	0
EgoCom [108]	2021	No	No	34	1	1	1	2	1	0	175	13.22	38.5	38.5	0	0	0	38.55
EGO-CH [117]	2020	Yes	No	70	2	1	1	4	1	0	180	12.62	29.14	0	0	0	0	0
EPIC-KITCHENS-100 [27]	2020	Yes	Yes	37	45	3	1	6	3	0	700	17.15	100	100	0	0	0	0
LEMMA [63]	2020	No	No	8	7	-	1	2	1	7	445	2	10.1	0	0	0	0	0
Charades-Ego [129]	2018	No	Yes	112	112	4	1	1	112	0	4000	0.52	34.4	0	0	0	0	0
EGTEA Gaze+ [79]	2018	No	No	32	1	1	1	2	1	0	86	18.76	28	28	28	0	0	0
EgoHands [8]	2015	No	No	4	3	1	1	2	1	0	48	1.5	1.2	0	0	0	0	1.2
BEOID [29]	2014	No	No	5	6	1	1	1	1	6	58	1	1	1	1	0	0	0
ADL [113]	2012	No	Yes	20	20	1	1	2	1	0	20	30	10	10	0	0	0	0
Disney [43]	2012	Yes	No	6	1	1	1	1	1	0	113	22.3	42	42	0	0	0	0
CMU Kitchen [72]	2009	No	No	55	1	1	1	1	1	0	175	8.28	24.16	24.16	0	0	24.16	0

cities). The Ego4D annotations are also of unprecedented scale and depth, with millions of annotations supporting multiple complex tasks—the result of 250K person-hours of annotation effort. 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.

Table I compares Ego4D (both the original v1 release and the 2023 v2 release) with the most relevant publicly available datasets of egocentric videos according to different dimensions, including whether the videos are unscripted (Uscr.), if the subjects operated in native environments (Nat. Env.), the number of subjects (Subj.), locations (Loc.), countries (Cntr.) in which the data has been collected, scenarios (Scen.), tasks (Tsk.), devices used for the data collection (Dev.), 3D scans of the environments (3D Env.), videos (Vids), average video lengths in minutes (Avg. Len), as well as the total number of hours, hours of video containing audio, gaze, stereo video, IMU measurements, and synchronized videos from multiple subjects. As can be observed in the table, Ego4D offers a unique variety in terms of subjects (931) locations (74), countries (9), scenarios (136), tasks (16) and devices used for the data collection (14). Distinctively from other datasets, Ego4D contains 15 3D scans of the environments in which the data has been collected, a large number (9,645) of untrimmed (average length of 24.11 minutes) videos and significant numbers of hours of overall video (3,670), video containing audio (2,533.16), gaze (33.35), stereo video recordings (79.82), IMU measurements (866.22) and synchronized videos from multiple subjects (497.99).

III. COLLECTING EGO4D: WHAT, WHERE, WHO

Next we overview the dataset and how we created it. The data and annotations are publicly available under an Ego4D license.

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 consortium consists of 14 teams from universities and labs in 9 countries and 5 continents (see map in Fig. 1). Each team recruited between 10-100 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 v1).

In the following, we overview the consortium's approach to recruiting and training diverse participants (Section III-A), the cameras and modalities recorded (Section III-B), the span of scenarios captured (Section III-C), and protocols to ensure responsible collection (Section III-D).

A. Participants

Next we describe the "who" of Ego4D: the camera wearers.

Recruitment: The consortium adopted three main strategies to recruit diverse participants. Most universities (Georgia Tech, Indiana U., U. of Minnesota, National University of Singapore, KAUST and U. of Bristol) recruited through local or national social media adverts, choosing the most popular platforms within their context. Others recruited participants through agencies (IIIT Hyderabad, U. of Tokyo, FRL, CMU and CMU Africa) where these agencies guaranteed the required diversity. A few universities (U. of Catania and KAUST) used word of mouth through recommendations from within the university community to recruit participants. Overall, diversity of participants was key while focusing on specific constraints—for example, group recruitment could only be done within households in certain countries due to COVID protocols.

Additionally, a mix of individual and group scenarios were captured. These varied from pairs of individuals to groups of six playing games or just engaging in a social gathering. Prior familiarity between the group is necessary for natural interactions, and thus participants were invited in groups (e.g. Georgia Tech) or tasked with recruiting other members of their household or from their acquaintances (e.g. U. of Bristol).

Camera Delivery and Training: Due to COVID restrictions, many universities could not invite participants on campus or meet with participants to deliver cameras and train on their usage. A variety of solutions were thus devised including mailing cameras (IIIT Hyderabad, U. of Bristol, U. of Indiana, and U. of Minnesota), with prepaid return postage at times. Remote guidance on camera usage was also provided via video conferencing platforms. These meetings were used to answer any questions the participants might have. Recordings were returned on harddisks (e.g. IITH Hyderabad), via secure cloud storage (e.g. U. of Bristol) or by returning the cameras (e.g. U. of Catania).

Diversity: 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

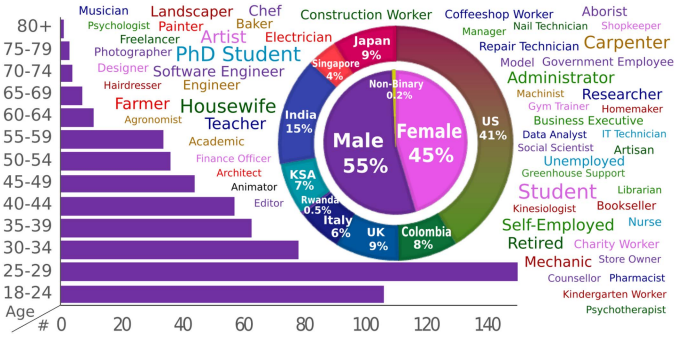


Fig. 3. Ego4D camera wearer demographics—age, gender, countries of residence, and occupations (self-reported). Font size reflects relative frequency of the occupation.

Fig. 3.[1] Participants 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.

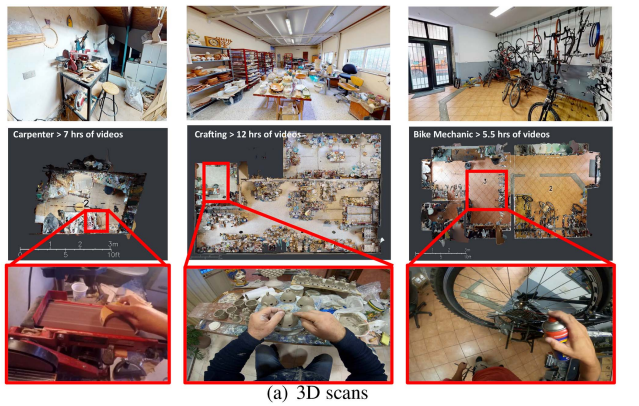
Participating camera-wearers reported working in seventy-nine different occupations, ranging from business executive to gym trainer to farmer, baker, electrician, homemaker, and beyond. Twelve universities collected self-reported data on camera-wearer country of birth or ethnicity (see Appendix of [57]). Differences in the granularity of this data prevent a collection-wide analysis, but the universities' efforts to recruit diverse participants are evident. For example, the 66 camera-wearers collecting data for KAUST originated from 20 different countries of birth. Similarly, 586 Ego4D camera-wearers volunteered to disclose their ethnicity, including Black, African or Caribbean, Hispanic or Latino, Asian, and Caucasian, among others. Sharing demographic information on camera-wearers contributing to an egocentric dataset is unique and marks an important step toward greater transparency.

Feedback From Camera Wearers: We invited participants to one-on-one interviews, following their experience with collecting videos for Ego4D. We recorded in-depth interviews with 16 participants, worldwide. Each answered 13 questions summarizing why they volunteered to collect data, how they selected the activities to record, their experience with the project and the hardware used, and personal reactions to watching their recordings.[2] Common reasons to participate include "passion for scientific research" and interest in wearable technology. Wearing a camera was deemed "a challenge at the beginning", "unusual", "awkward" but also "not uncomfortable", "did not distract from [social] interactions" and a "unique experience". Some expressed performance anxiety, particularly when capturing themselves performing their professional work. Watching the footage post recording was referred to as "surreal", "cool", "interesting", a reminder of "first-person video games" and "experiencing moments of life again".

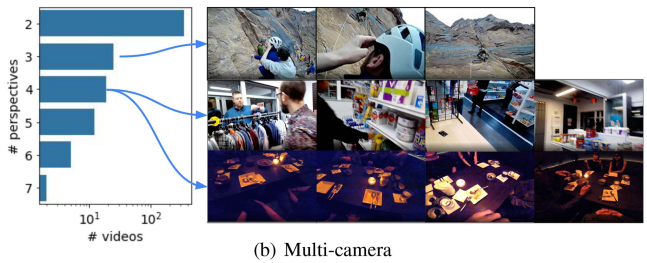
B. Cameras and Modalities

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, ORDRO 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. 4, 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. 4, left).

Fig. 4. Different perspectives between Vuzix (left) and GoPro (right).



(a) 3D scans



(b) Multi-camera

Fig. 5. (a) Some videos (bottom) have coupled 3D meshes (top) from Matterport3D scanners, allowing one to relate the dynamic video to the static 3D environment (middle). (b) Multi-camera data has multiple wearers simultaneously wearing cameras, giving multiple perspectives of the same event and facilitating cross-view video understanding and social interaction research.

In addition to video, portions of Ego4D offer several other data modalities: 3D scans,[3] audio, gaze,[4] stereo, multiple synchronized wearable cameras, and textual narrations. See Table II.

Each modality can support new research challenges. For example, having Matterport3D scans of the environment coupled with ego-video clips (Fig. 5(a)) offers a unique opportunity

[1]for 64% of all participants; missing demographics are due to protocols or participants opting out of answering specific questions.

[2]Short excerpts are available at https://youtu.be/5yXBcGCkgrA

[3]Two consortium members captured 3D scans of indoor environments using Matterport3D scanners: FRL captured 3D scans of three apartments, while University of Catania captured 3D scans of workshops and bakeries where hours of recordings were captured and registered to these scans. The scan was collected by the research team, as these are highly technical tasks that cannot be easily carried out by the participants.

[4]Eye trackers were deployed by Indiana U. and Georgia Tech only.

TABLE II
MODALITIES OF DATA IN EGO4D AND THEIR AMOUNTS

Modality:	RGB video	Text narrations	Features	Audio	Faces	3D scans	Stereo	Gaze	IMU	Multi-cam
V1 (hours):	3,670	3,670	3,670	2,535	612	491	80	45	836	224
V2 (hours):	3,891	3,891	3,891	2,548	612	491	80	45	836	227

"Narrations" are dense, timestamped descriptions of camera wearer activity (cf. Section IV). "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 [46] and Omnivore [56] video features.

for understanding dynamic activities in a persistent 3D context, as we exploit in the Episodic Memory benchmark (see Section VI-A). Multiple synchronized egocentric video streams allow accounting for the first and second-person view in social interactions (see Fig. 5(b)) and performing cross-view video understanding. Audio allows analysis of conversation and acoustic scenes and events.

C. Scenarios

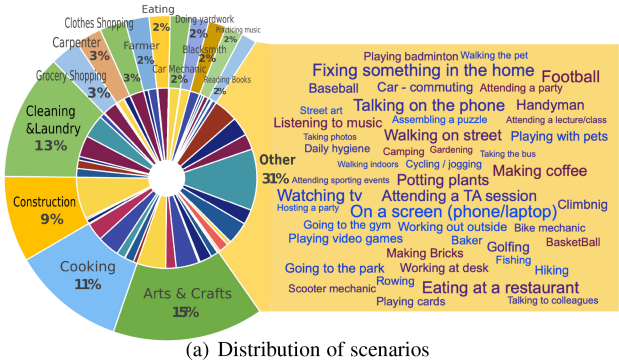
What activities belong in an egocentric video dataset? Our research is motivated by problems in robotics and augmented reality, where vision systems will encounter *daily life scenarios*. Hence, we consulted a survey from the U.S. Bureau of Labor Statistics[5] that captures how people spend the bulk 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 in our dataset lasts 20 minutes to an hour—significantly longer than the 10s clips often studied in third-person video understanding [65]. We find that after an initial period of 5-10 minutes, most camera wearers begin to forget they are wearing the camera and engage naturally with their environment. In this way, we capture unscripted activity while being mindful of the scenarios' coverage.

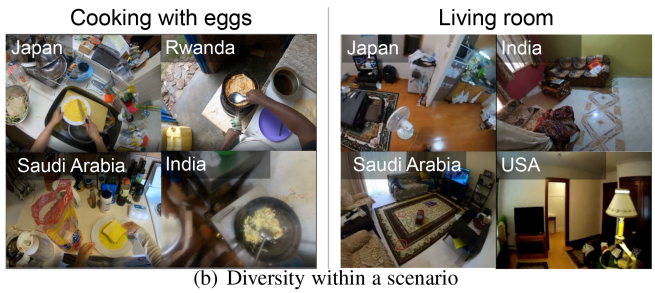
The exception is for certain multi-person scenarios, where we asked participants at five sites who had consented to share their conversation audio and unblurred faces to take part in social activities, such as playing games. We leverage this portion of Ego4D for the Audio-Visual and Social Interaction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

Participants recorded both indoor and outdoor scenarios, as well as recordings in their leisure time and while working. Of particular notice are recordings from CMU and Catania where specialist workers (e.g. farmers, bakers and construction workers) were capturing their days of work, showcasing high levels of skill.

Fig. 6(a) 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. Thanks to Ego4D's geographic diversity, there is substantial visual variety even within the same scenes and activities. Fig. 6(b) shows examples of participants cooking with eggs in four countries (left), and examples of living rooms across the world. These examples highlight how different the same activity or scene can appear in different locations due to cultural differences, availability of equipment, and personal preferences. Overall, the 931 camera wearers bestow Ego4D with a glimpse of daily life activity around the world.



(a) Distribution of scenarios



(b) Diversity within a scenario

Fig. 6. Scenarios in Ego4D. (a) 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). (b) Within a given scenario, there is tremendous visual diversity, e.g., as seen here by people cooking eggs (left) or using a living room (right) in different parts of the world.

D. Responsible Data Collection

From the onset, privacy and ethics standards were critical to this data collection effort. While necessary for any video collection, the first-person daily life nature of Ego4D accentuates such considerations.

[5][Online]. Available: https://www.bls.gov/news.release/atus.nr0.htm

1) Standards of Responsible Collection: Each partner was responsible for developing a policy for responsible data collection. In particular, all members of the consortium prepared individual applications for their Institution Review (or Ethics) Boards. These applications covered the goals of the consortium and the protocols to be followed. They varied per institution and were reviewed and approved individually. The leads provided clarifications and made amendments based on the recommendations of these boards. Recruitment and recordings only started after these applications were approved.

While specifics vary per site, there is a set of standards generally adopted across the dataset:

- Comply with own institutional research policy, e.g., independent ethics committee review where relevant
- 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[6]
- 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, (such that faces are preserved in the frames and audio can be maintained), or else in public spaces where faces and other PII are blurred.

2) Review, De-Identification, and Management: A few universities required participants to review the footage themselves so as to approve the recordings (e.g. University of Bristol, Georgia Tech, and Indiana University). This was followed by manual inspection of the recordings, whether by the researchers themselves watching the footage or by hired administrators.

For any video without explicit informed consent that contains PII, the university owning that data performed de-identification before releasing it in Ego4D. That includes indoor settings with multiple participants present, PII captured accidentally such as an address on an envelope or a reflection of the wearer's face on a mirror or a surface, as well as videos recorded outdoors in a public space where bystanders or cars appear in the footage.[7]

These videos were de-identified using advanced video redaction software, open source tools, and hours of human reviews. Each university partner undertook the de-identification effort for their own data. Typically, this started with an automated process to detect faces and license plates and blur them, followed by manual review of all outputs from automated blurring. Other PII data such as written names/addresses, phone screens/passwords or tattoos had to be manually identified and blurred per-frame. The time required to de-identify a video ranged from 1.5× to 10× the length of the video, with high variance depending on the scenario and content.

Recording large-scale videos requires careful data management, backup, and anonymizing the link between the recording and any collected consent forms or communication with the participants. All members of the consortium took particular care in safely maintaning the information for any future communication with participants. This information is only stored locally, and safely, with access for approved need-to-know individuals. In case where GDPR-compliance is required, the ownership of the data remains with the participants themselves, and thus the link between the participants and their recordings should be maintained. In other cases where the ownership of the data was transferred the consortium member, consent forms were retained without a direct link to the person's recordings, establishing complete anonymity.

3) Informed Consent and Acceptable use: Informed consent is an important aspect of data collection from human subjects. When data is collected for the purpose of training deep models, it can be difficult for participants to understand what they are consenting to exactly. Teams addressed this issue by giving examples from domains like autonomous driving to illustrate what deep models are and how they use data. A more subtle issue concerns the researchers downloading and using the data. Since the data was collected under an IRB or its equivalent, the researchers collecting the data would typically have completed required ethics training (e.g. CITI training in the U.S.) This training is valuable, as it can sensitize researchers to the concerns that arise in working with protected data. However, researchers who download and use the data may not have received training in human subjects protections. The Ego4D license and data use agreement helps to address this by identifying prohibited activities. As egocentric vision fundamentally depends on data collected from human subjects, this is another reason for including ethics training as a standard part of graduate student curriculla.

4) Known Biases in the Dataset: While Ego4D pushes the envelope on 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 workers in two sites in Africa. This means that there will be at least subtle ways in which the language-based narrations (discussed next) are biased towards their local word choices.

5) Accessibility of the Dataset: At 3,670 hours of video, we are mindful that Ego4D's scale can be an obstacle for accessibility for some researchers, depending on their storage and compute resources. To mitigate this, we have taken several measures.

First, we provide precomputed action features (SlowFast 8x8 with ResNet 101 backbone pretrained for Kinetics 400), an optional starting point for any downstream work. Second, only portions of the data constitute the formal challenge train/test sets for each benchmark—not all 3,670 hours (see Table IX). As Ego4D annotations increase, we will create standardized mini-sets. Third, we provide the option to download only the data targeting an individual benchmark or modality of interest.

[6]Only video of this type is used in our Audio-Visual Diarization and Social Interaction benchmarks.

[7]Note that due to differences in approved IRBs, some outdoor recordings of passers by were deemed acceptable in public spaces where there was no manipulation or direct interaction with study personnel, while other universities blurred all identities of incidental capture.

Fig. 7. *Example narrations at keyframes of video.* #C refers to the camera-wearer. The last row shows narrations that include other people that participate in activities with the camera-wearer (denoted by other letter tags, e.g., #O, #X).

Our visualization tool[8] allows researchers to browse the video and annotations and select subsets most relevant to their own research, without downloading in its entirety. Finally, our team directly assists users to promote accessibility, such as tutorials on dataset access at the CVPR workshops, a YouTube channel with resources and trainings,[9] office hours on Zoom, and a forum for responding to questions.[10]

We observe that thus far academic teams actually dominate the participation in Ego4D's formal challenges. For example, at ECCV 2022, 80% of the teams were comprised of majority academic contributors, while 20% have majority authors from industry. This is a positive sign about the accessibility of the dataset, since academic groups typically have much more modest computational resources than industry labs.

IV. Narrations of Camera Wearer Activity

Before any other annotation occurs, we pass all video through a *narration* procedure which attaches free-form natural language descriptions of the camera wearer activity at frequent time points in all the video (see Fig. 7). The narrations are a form of "pre-annotation". They are not the ground truth for any particular benchmark. Rather, the narrations allow us to (1) perform text mining for data-driven taxonomy construction for actions and objects (see Section VII-B), (2) sort the videos by their content to map them to relevant benchmarks, (3) identify temporal windows where certain annotations should be seeded, and (4) browse the video content at scale (see Section V).

Narration Procedure: Inspired by the pause-and-talk narrator [28], annotators are asked to watch a 5 minute clip of video, summarize it with 1-3 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. Each video receives two independent narrations from different annotators.

Specifically, narrators are provided the following prompt: *"Pretend as you watch this video that you are also talking to a friend on the phone, and you need to describe to your friend everything that is happening in the video. Your friend cannot see the video."* This prompt is intended to elicit detailed descriptions that provide a play-by-play of the action. Each narration thus corresponds to a single, atomic action or object interaction that the camera wearer performs (e.g., "#C opens the washing-machine" or "#C picks up the detergent", where the tag #C denotes the camera wearer). Importantly, our narrations also capture interactions between the camera-wearer and others in the scene, denoted by other letter tags, e.g. #X (e.g. "#C checks mobile while #X drives the car", "#C passes a card to #Y"). We also ask the annotators to provide a short summary description of the entire clip (indicated with #summary in the dataset). See Fig. 7 for example narrations.

Narration Analysis: 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).

Fig. 8 (top row, left) shows the distribution of frequency of narrations across all videos in the dataset. Depending on the activities depicted, videos are annotated at varying frequencies. For example, a video of a person watching television is sparsely annotated as very few activities occur (0.17 sentences/minute), while a video of a person harvesting crops, performing repetitive actions is densely annotated (63.6 sentences/minute). Fig. 8 (top row, middle and right) show the distribution of length of the collected narrations. The individual timepoint narrations are short, highlight a single action or object interaction, and have an average of 7.4 words. Though short, these narrations cover a variety of activities ranging from object interactions, tool use, camera wearer motions, activities of other people etc. In contrast, the summary narrations are longer (on average, 16.8 words) and describe activities at a higher level (see rightmost histogram).

Finally, we study the diversity of the video dataset by looking at the frequency of occurrence of words in the narrations collected for videos of each scenario type. Fig. 8 (bottom) shows word clouds depicting objects that prominently feature in videos across various scenarios. The word clouds highlight characteristic objects per scenario (e.g., bowl, spoon, plate in "Cooking" videos; card, dice, pawn in "Playing board games" videos) while also hinting at common objects across all scenarios

[8][Online]. Available: https://visualize.ego4d-data.org/
[9][Online]. Available: https://www.youtube.com/@ego4d954
[10][Online]. Available: https://discuss.ego4d-data.org/

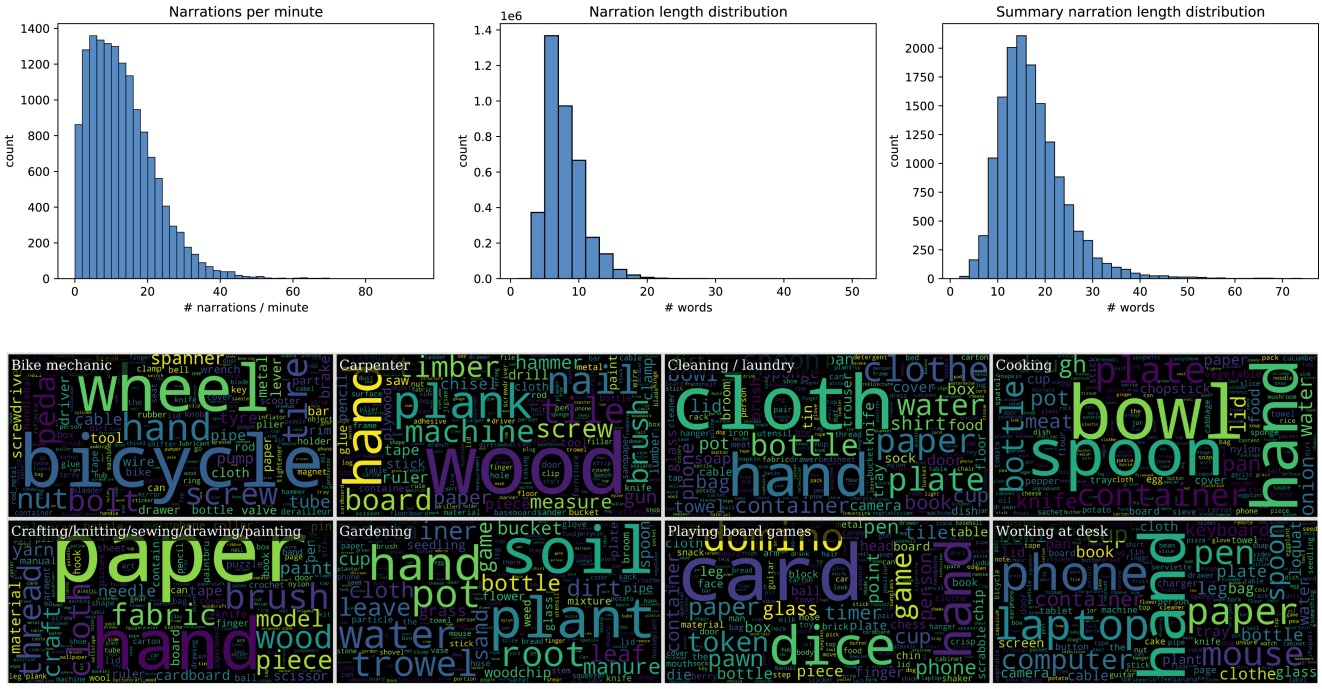


Fig. 8. *Narration analysis.* Top row: Distribution of narration frequency (left) and the distribution of narration and summary lengths (middle and right). Summaries are naturally longer, and describe activities at a higher level compared to individual action narrations. Bottom row: Diversity in the distribution of objects across eight common scenarios extracted from narrations.

(e.g., hands, paper, phones). The diversity in narrations collected highlights the diversity of video content captured in the dataset.

Narrations' Impact: To our knowledge, ours is the largest repository of aligned language and video (e.g., HowTo100M [95], an existing Internet repository with narrations, contains noisy spoken narrations that only sometimes comment on the activities taking place).

Beyond our uses of narrations to index the data and form taxonomies, as described above, the narrations are themselves a contribution of Ego4D, potentially valuable for research on video with weakly aligned natural language. In fact, since its release, the scale, diversity and strong temporal alignment in Ego4D narrations has enabled the training of state-of-the-art video-language representations. For example, these representations have been the basis of challenge-winning egocentric video models on several Ego4D benchmark tasks [82] as well as data-efficient imitation learning approaches for real-world robot learning tasks [103].

V. Visualizing Ego4D Data

The scenarios overviewed in Fig. 6 provide rough guidelines for what we expect to see in the dataset. However, the unscripted nature and wide variety of participant behaviors make the collection of 3,670 of video somewhat opaque without further mining and analysis. To that end, we built a visualization tool that allows users to browse our dataset, visualize annotations relevant to all our benchmark tasks, and perform semantic search to retrieve videos from the dataset using natural language queries. The tool is hosted at https://visualize.ego4d-data.org/. Next, we visualize and analyze the videos for their objects, scenes, and activities to gain a deeper understanding of what they contain.

A. Objects

We start by analyzing objects in videos. The right part of Fig. 18 (discussed later in Section VII-B) shows the top 45 (out of 478) object categories identified from the video narrations. In terms of coverage, just 11.6% of the identified objects account for over 41.8% of the classes in COCO [85], a standard object detection dataset. The remaining 88% represents new object classes from diverse scenarios (see word clouds in the bottom of Fig. 8 for examples). Further, these objects are captured in non-canonical viewpoints in the egocentric view, and pose a significant challenge for detection models. A Faster-RCNN [121] model trained to detect objects achieves just 20.1 mAP (mean average precision) over the 51 overlapping classes, compared to 39.8 mAP on COCO.

B. Scenes

To analyze the diversity of scenes in the dataset, we generate scene scores using a pre-trained scene classification model [154] over a subset of 1.4k hours of video. Specifically, we assign scores to each frame of a video, and average scores across all frames to arrive at a scene label per video. Note that these labels are approximate as the scene classification models were trained on a dataset of canonical third-person scene images (not egocentric video frames). Nevertheless, they offer insights into the prevalent scenes in our dataset. We visualize the distribution of these labels in Fig. 10 (left) and over the scene hierarchy from [154] in Fig. 10 (right). A majority of Ego4D is collected in indoor environments that humans typically use, however a large proportion of videos (20%) are in outdoor scenes ranging from water (fishing, kayaking, scuba diving) to forests (foraging, hiking) to urban areas (daily commute). Together, the scenes

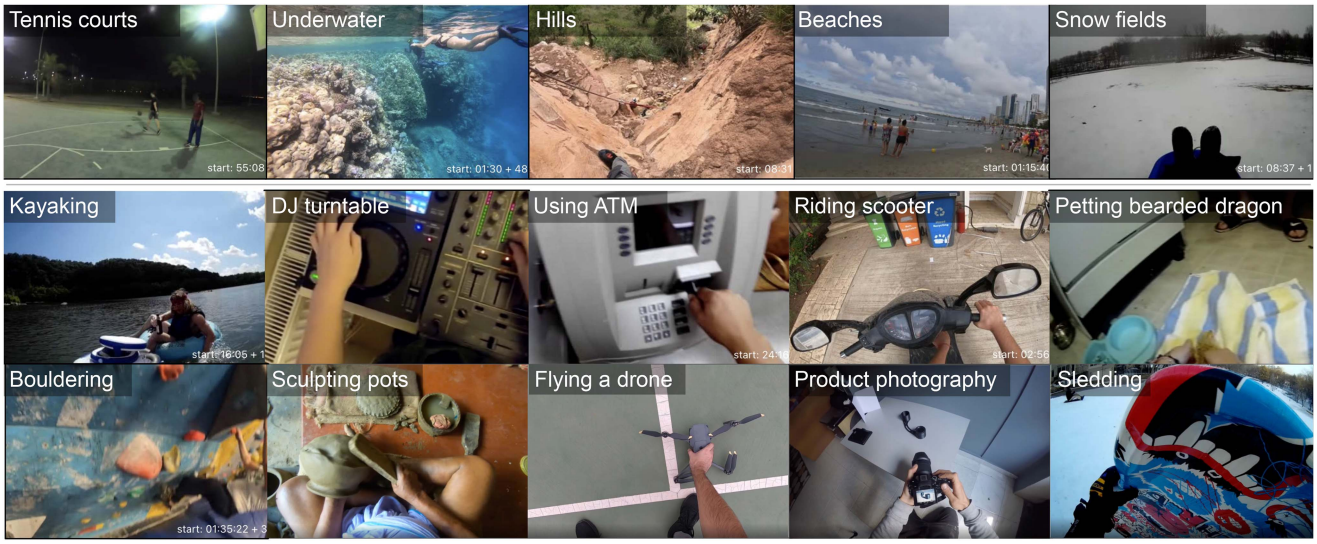


Fig. 9. *Diversity in scenes and activities.* Top: Relatively uncommon scenes in the scene hierarchy. Bottom: Unique activities captured by clustering narrations and visual content.

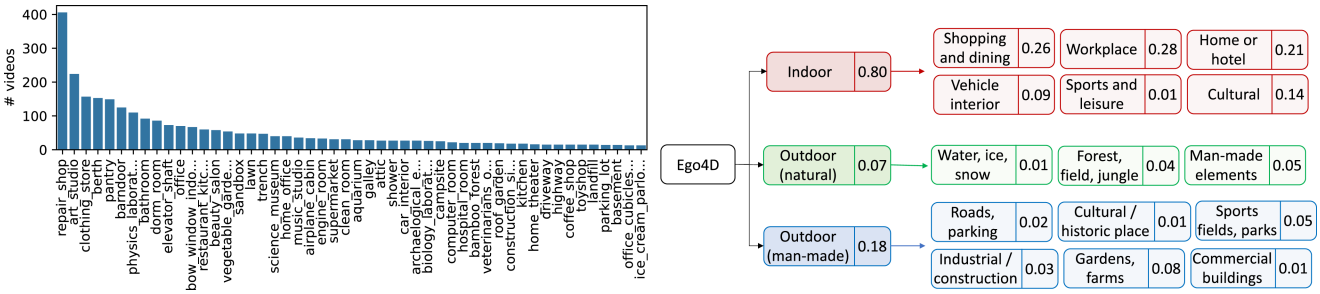


Fig. 10. *Scene diversity*. Left: Top 50 scenes across videos in the dataset as computed by a pre-trained scene classification model. Right: Distribution of scenes over the Places365 hierarchy [154]. Numbers represent the fraction of videos belonging to the scene hierarchy node.

capture human activity in diverse locations. Some examples of uncommon scenes are shown in Fig. 9 (top).

C. Activities

Next, we investigate activities in the dataset along several axes beyond the scenario-level labels in Section III-C.

Activity Clusters: We cluster video segments based on their visual features and text embeddings of paired narration summaries. While the large clusters are typically covered by the scenario labels, we identify several other clusters with unique visual content. These include activities like kayaking, using an ATM, petting a bearded dragon, rock climbing, riding a scooter, using a DJ turntable. Fig. 9 (bottom) shows samples of several such activities.

Repetitiveness in Activities: We estimate how much the visual features across the video changes over time as a metric of repetitiveness in visual content. Specifically, we compute the the standard deviation of Omnivore [56] features over a rolling-window across the video. A high standard deviation implies more visual change. The distribution is shown in Fig. 11,

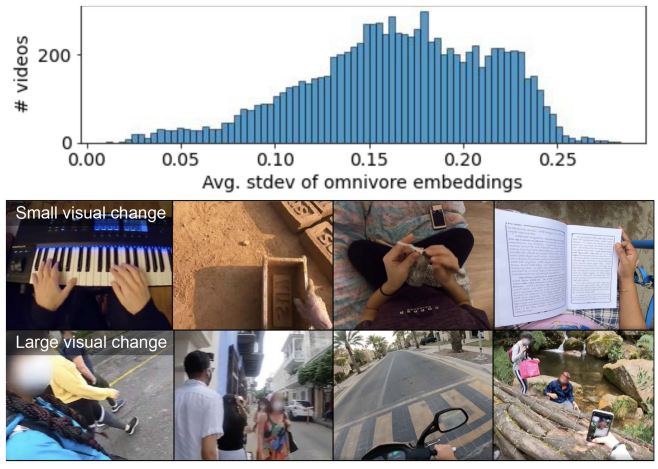


Fig. 11. *Distribution of repetitiveness of activities*. A high standard deviation suggests substantial viewpoint and scene changes (bottom row). A low standard deviation implies lower visual changes across the sequence (e.g., playing piano, making bricks, reading in top row), but does not imply less "interesting" content. See text.

Fig. 12. The Ego4D benchmark suite centers around the first-person visual experience—from remembering the past, to analyzing the present, to anticipating the future. The supplementary video available here https://ego4d-data.org/ overviews each task.

with examples alongside. Note that a low visual diversity does not imply that the activity is less "interesting". For example, a video of a person playing the piano has low visual diversity, but it captures a complex, skilled activity that is relevant to our benchmarks.

VI. Benchmark Tasks

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 Fig. 12. The following sections introduce each task and its annotations. The dataset release has annotations for 48-865 hours of data per benchmark, on top of the 3,670 hours of data that is narrated. See Table IX.

We developed baseline models drawing on state-of-the-art components from the literature in order to test drive all Ego4D benchmarks. At the time of writing, we have run two formal Ego4D competitions (at CVPR 2022 and ECCV 2022) where we invited the research community to make progress on these challenges. The sections below capture the current state of the art, which has already exceeded our originally published baselines, as the result of exciting and substantial new efforts in the field leveraging Ego4D.

A. Episodic Memory

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 [137]: 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.

1) Task Definition and Annotations: Given an egocentric video and a query, the Ego4D Episodic Memory task requires localizing the answer in the user's past video. We consider three query types—natural language queries, visual queries, and moments queries—overviewed in Fig. 13 and explained next. Unlike traditional Q&A systems, the response is not a natural language answer. Rather, for all query types, the output is video localization—temporal and optionally spatial (in the frame or in the 3D environment).

TABLE III
The NLQ Templates Capture a Diverse Set of Queries That People Might Ask to Augment Their Memory and Recollect Objects, Places, and People in Their Everyday Experience

Category	Template
Objects	Where is object X before / after event Y? Where is object X? What did I put in X? How many X's? (quantity question) What X did I Y? In what location did I see object X ? What X is Y? State of an object Where is my object X?
Place	Where did I put X?
People	Who did I interact with when I did activity X? Who did I talk to in location X? When did I interact with person with role X?

For ***natural language queries*** (NLQ), 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 easily deducible. Queries can be related to objects, places, people, and activities that appeared in the episodic memory of the camera wearer, and should not require an external knowledge base to answer. NLQ is a challenging multimodal task requiring both visual (recognizing events, objects, object states, places) and linguistic (breaking down reasoning, understanding relations) understanding. Concretely, given an egocentric video $\mathcal{V}$ and a natural language query $\mathcal{Q}$, the goal is to identify a 'response window' of temporally contiguous frames r in $\mathcal{V}$, such that the answer to $\mathcal{Q}$ can be deduced from r.

To generate an annotated NLQ dataset for interesting queries, we first use the narrations (cf. Section IV) select a pool of 227 hours of video more likely to contain a variety of objects, non-repetitive actions, and movement by the camera wearer between multiple rooms (based on the entropy over a set of manually curated navigation verbs). Each clip is from 8 to 20 minutes long. Then we devise a set of 13 template questions (see Table III) meant to span things a user might ask to augment their memory,

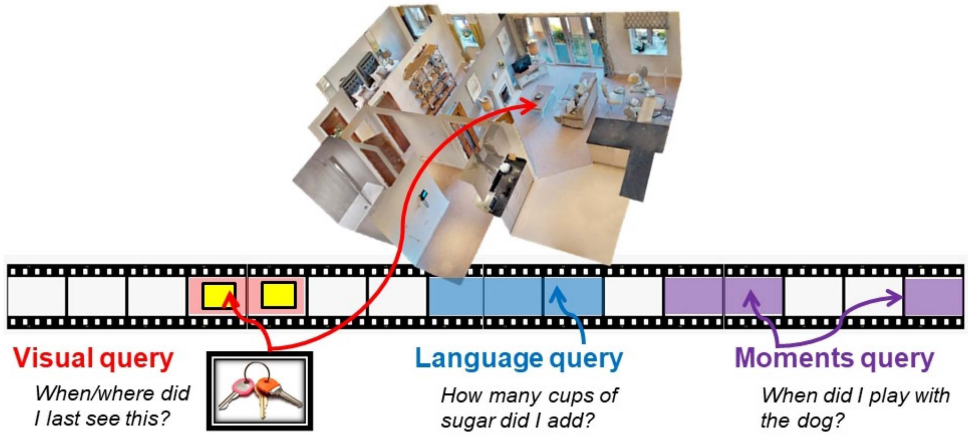


Fig. 13. Episodic Memory's three query types. Visual queries (VQ) ask about the location (in the video and optionally 3D space) where an object depicted in an image was last seen. Natural language queries (NLQ) ask free-form natural language questions about the happenings in the video. Moments queries (MQ) ask for temporal localization of all the instances of a given activity type, for a taxonomy of more than 100 activities.

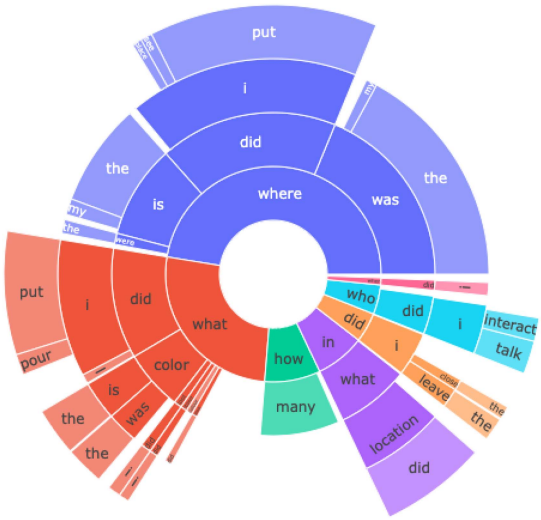


Fig. 14. Distribution of query words in NLQ. Center ring shows first word, followed by outer rings showing subsequent words.

such as *"what is the state of object X?"*, e.g., "did I leave the window open?", or *"where is object X before / after event Y?"*, e.g., "where was the dog before I opened the door?". Annotators express queries in free-form natural language, and also provide the slot filling (e.g., X = dog in "where was the dog?"). In this way, the templates focus the type of questions, but the actual data contains free-form language created by the annotators. Fig. 14 shows the distribution of initial query words in the sentences they generated.

For ***visual queries*** (VQ), the query is a static image of an object, which, importantly, is taken from a frame disjoint from the input video. 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.[11] 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]?"). By enabling visual queries, as opposed to categorical queries, this is a form of open-world object localization.

Given an egocentric video $\mathcal{V}$, a query object o specified via a static visual crop v, and a query frame q, the goal is to identify when the object o was last seen in the video before the query frame q. For the 2D case (VQ2D), the response is again specified as a response track, but here in terms of a contiguous series of 2D bounding boxes surrounding the object o in each frame: $r = \{r_s, r_{s+1}, \ldots, r_{e-1}, r_e\}$, where s is the frame where the object o (at least partially) enters the camera-wearer's field of view, e is the frame where the object exits the camera-wearer's field of view, and r_i is a bounding box (x, y, w, h) in frame i. If the object appears multiple times in the video, the response only refers to the 'most recent occurrence' of the object in the past. This reflects the use case where the user is interested in the most recent state of the depicted object. See Fig. 15, top. For the 3D case (VQ3D), when a 3D scan of the environment associated with the video is available, the response additionally includes a 3D displacement vector $\Delta d = (\Delta x, \Delta y, \Delta z)$ between the 3D location where the query was made (i.e., at query frame q), and the 3D location in the environment where the object was last seen (i.e., at the end of the response track r_e). That 3D location was obtained by annotators drawing a 3D bounding box in the environment scan for the object's last position. See Fig. 15, bottom.

For VQ, we again use the narrations to prioritize labeling videos where the camera wearer moves around in the environment, which makes queries more challenging. In addition, we select all videos for which we have an accompanying 3D scan available. In total our benchmark offers 262 total hours of VQ2D queries from hundreds of environments and 19 hours of video

[11] We provide 3D response ground truth for those videos that have an accompanying Matterport3D scan of the surrounding environment.

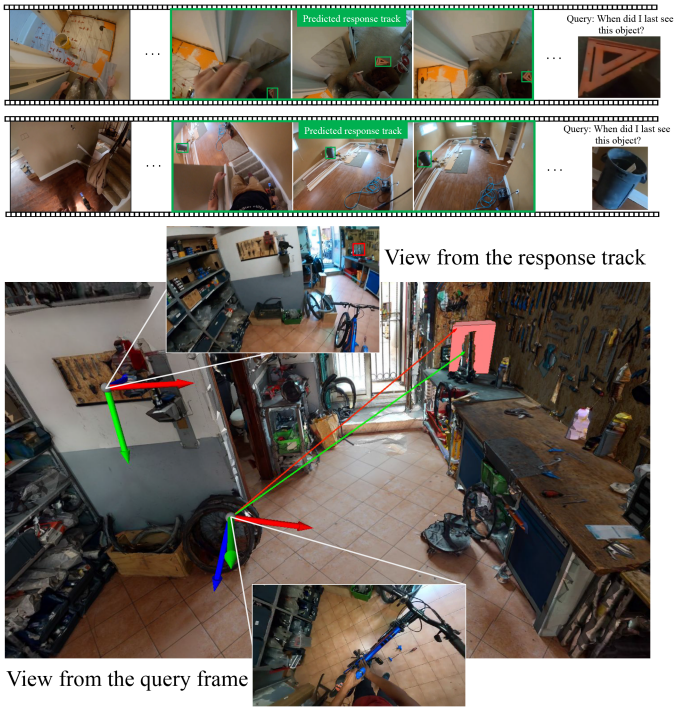


Fig. 15. Visual queries in 2D (top) and 3D (bottom). Top: Two example VQ2D localization. We show the visual crop of the query on the right, and the predicted response track in the center (3 uniformly sampled images). The model localizes and tracks the most recent occurrence of the object—note the query and localized instances are non-identical. Bottom: Top image shows view from the last frame of the response track with the target object annotated with a 2D red bounding box. Bottom shows the view from the query frame. The target object is annotated with a 3D red bounding box at the top right of the figure. The figure shows the ground-truth (green) and the predicted (red) 3D displacement vectors.

TABLE IV
EPISODIC MEMORY BENCHMARK RESULTS: WE REPORT THE VALIDATION AND TEST RESULTS ON THE FOUR EPISODIC MEMORY TASKS

Visual Queries 2D Localization	stAP @ IoU=0.25 ↑		tAP @ IoU=0.25 ↑	
Method	Val	Test	Val	Test
Winner [145]	20.0	18.0	27.0	26.0
Ego4D Baseline	12.0	13.0	20.0	21.0

Visual Queries 3D Localization	Overall Success ↑		QwP ↑	
Method	Val	Test	Val	Test
Winner [87]	-	26.0	-	66.0
Ego4D Baseline	1.2	8.0	1.8	16.0

Natural Language Queries	Mean recall@1 ↑		recall @5 IoU=0.3 ↑	
Method	Val	Test	Val	Test
Winner [19]	12.9	13.2	24.8	22.9
Ego4D Baseline	4.2	4.0	10.7	8.8

Moment Queries	average mAP ↑		recall@1 IoU=0.5 ↑	
Method	Val	Test	Val	Test
Winner [19]	23.2	23.6	40.4	41.1
Ego4D Baseline	6.0	5.6	25.2	24.2

For each case, we report the winning entry from our recently concluded Ego4D challenge held at ECCV 2022 [1], and our baselines [57]. The primary challenge metrics are shown in the second column of each table.

from 4 unique environments for VQ3D queries. Clips vary in length from 5 to 16 minutes each.

For ***moments queries*** (MQ), the query is the name of a high-level activity or "moment", and the response consists of all temporal windows where the activity occurs. Specifically, MQ poses the following request: 'Retrieve all the moments that I do X in the video', e.g., "When did I read to my children? When are all the times I exercised?" Compared to the free-form natural language queries, these queries all center around the first-person's activity, and the target 'X' is from a pre-defined taxonomy of action categories, such as 'interact with someone' or 'use phone'. Moments aim to 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 (Section VI-E).

Given an egocentric video $\mathcal{V}$, and a query action category c, the MQ goal is to retrieve all the instances of this action category in the video. The response is a set of action instances of the category c $\Phi_c = \{\phi_n = (t_{n,s}, t_{n,e}, s_n)\}_{n=1}^{N}$, where n is the number of instances for this category, $t_{n,s}$ and $t_{n,e}$ are start time and end time of the n^{th} instance respectively, and s_n is its prediction confidence.

For moments, we established a taxonomy of 110 activities in a data-driven, semi-automatic manner by mining the narration summaries—one taxonomy for each scenario of interest (cooking, cleaning, shopping, handyman, farmer/gardner). See Section VII-B for details and Fig. 16 for the resulting moments taxonomy. We ask 3 annotators to label each 8 minute clip with each and every temporal segment containing a moment instance, given the taxonomy. We take the union of their labels, following [26]. The average duration of an instance is 45 seconds, and most clips have 1-20 instances. In total, the v1 release has ∼74K total episodic queries spanning more than 800 hours of video.

2) Relation to Existing Work: Episodic Memory has some foundation in existing vision problems, but it 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 plus vision focuses on captioning and question answering for isolated instances of Internet data [70], [144], NLQ is motivated by queries about the camera wearer's own visual experience and operates over long-term observations. VQ upgrades object instance recognition [13] to deal with video (frequent FoV changes, objects entering/exiting the view) and to reason about objects in the context of a 3D environment. Finally, MQ can be seen as activity detection [84] but for the activities of the camera wearer.

3) Baselines and Evaluation Metrics: We developed baseline models to provide a starting point upon which future work can build. See [57] for details on the baseline models and architectures, which we only briefly review here. Table IV summarizes the current SoTA on all EM tasks. For each task, we present our baseline numbers alongside the current best reported in the literature, as judged in the recent ECCV 2022 competition where 16 teams competed on Episodic Memory tasks.

As NLQ baselines, we explore two prior methods: 2D temporal adjacent networks (2D-TAN) [150] and span-based localization networks (VSLNet) [148]. 2D-TAN takes adjacent moment candidates as the temporal context on a 2D temporal map and retrieves the most relevant moment from those candidates, as predicted by a CNN trained to predict the IoU with the ground truth response window. VSLNet treats the input untrimmed video as a text passage and uses a span-based approach to identify the relevant sections semantically related to the natural language query. Both the language and video features are

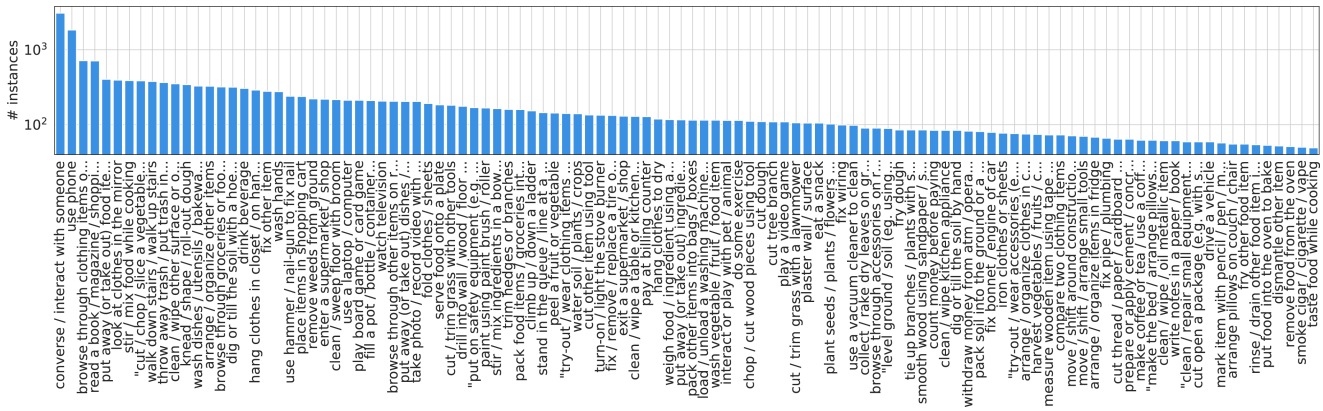


Fig. 16. *Distribution of moments labels* showing number of instances per 110 categories across 5 scenarios and 300 hours of video. Note that these are only the Moments for Episodic Memory with temporal window annotations in the current release; Ego4D has many other scenarios and activities not reflected in this distribution.

encoded with a Transformer. Then the encoded query is used to attend to the relevant parts of the video clip. NLQ's evaluation metric is the top-k recall at a given temporal intersection over union (tIoU) threshold.

As a VQ baseline, we address the VQ2D case with a mix of detection and tracking. Specifically, for VQ2D we perform frame-level detection for the query object using a variant of Faster-RCNN we call Siam-RCNN, which is trained to score proposals based on a Siamese head applied to the visual crop feature. Then we find the peak in detection scores over time and initialize a tracker forward and backward from that point to recover the response track. To further address the VQ3D case, we estimate camera poses for the input video frames, estimate the depth of the detected object produced by VQ2D , and retrieve its 3D position from the query frame. VQ adopts temporal and spatio-temporal localization metrics (tIOU and stIoU) as well as timeliness metrics that encourage speedy searches.

For the MQ baseline, we treat moment queries as a temporal action detection task followed by simple post-processing. We adopt VSGN [151] for temporal action detection, which extracts SlowFast features for each video snippets, and then feeds the features to a graph pyramid network that predicts the scores and refines the locations of the anchors. MQ adopts a popular metric used in temporal action detection: mAP at multiple tIoU thresholds, as well as top-kx recall.

4) Discussion: Episodic memory offers challenging multi-modal video search tasks with wide applications in helping users retrieve relevant pieces of their visual experience. The performance of the existing state-of-the-art video localization models highlight the needle-in-a-haystack nature of the tasks, where response windows occupy only a tiny fraction of the input video (e.g., a few seconds within 8-16 minutes of video). The variable types of query studied here—visual crops, free-form language, named activities—call for new flexible video search architectures and more powerful methods for relating the text and visual modalities. In addition, the long-form nature of the input video accentuates the need for scalable video models, and calls for advances in sub-linear time search to avoid scrutinizing every moment of every long video to find the answer. Our data and results also show the importance of tackling the long-tailed distribution of human activity. As an in-the-wild uncurated collection of first-person experience, Ego4D reveals the real-world setting is inherently class-imbalanced, reinforcing the need for continued work on low-shot learning, particularly for video understanding.

B. Hands and Objects

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.

1) Task Definitions: We first define an *object state change* in terms of attributes in three conceptual dimensions: time, space and semantic. Then based on these three dimensions we propose three tasks to evaluate a models ability to understand object state changes along the temporal, spatial and semantic dimensions.

In the temporal dimension, we consider an object state change to be represented by three distinct temporal points in the video. *(a) Point-of-no-return:* The point-of-no-return (PNR) is the frame I_{pnr} in a video that identifies the beginning of an object state change that cannot be easily reversed. *(b) Pre-condition:* The pre-condition is defined as some frame I_{pre} that marks a moment prior to the state-change in which the related objects were visible within the field of view of the camera. *(c) Post-condition:* The post-condition is some frame I_{post} at which the completion of the state change is visible after the point-of-no-return. Any model designed to recognize object state changes should be able to represent these three distinct temporal stages. See Fig. 17.

Fig. 17. Hands and Objects: Example object state changes defined by pre-condition, PNR, and post-condition frames.

In the spatial dimension an object state change can be represented by the bounding box of the object at the PNR, pre-condition and post-condition, along with any tools involved in performing the state change. Tools offer extended capabilities of the actor's hand, such as using an electric saw to cut a piece of wood in half. Any model designed to recognize an object state change should be able to identify bounding boxes (or segmentation masks) that capture the hands, tools and objects undergoing the state change.

In the semantic dimension we represent an object state change through the human action (verb), the object identity (noun) and the type of state change applied. The same state change can be performed on different objects using different tools. For example, cutting a piece of wood with electric saw and cutting a piece of paper with scissors are different interactions with different objects and different tools but they both result in the same object state change of *being cut*.

Based on the three dimensions of an object state change described above, we propose the following three tasks.

1) *Temporal. 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) *Spatial. 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) *Semantic. Object state change classification:* Given a short video clip, the goal is to classify whether an object state change has taken place or not.

2) Relation to Existing Work: Limited prior work considers object state change in photos or video; Ego4D is the first video benchmark dedicated to the task of understanding object state changes. The task is similar to action recognition 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 memorize specific action or object observations.

Existing approaches for modeling object states or their changes can be categorized into two research lines. The first deals with collections of images. A representative dataset for this purpose is the MIT States dataset [61]. By considering object states as object attributes (e.g., burnt, sliced), this line of work studies attribute-object composition, e.g., composition with context [97], modeling attributes as operators [101], and an architecture for compositional reasoning [114].

The second research line deals with video and views an action as a state transformation over time. One direction is the discovery of object states or manipulating actions. Fathi et al. [44] explore object state detection in video using a weakly supervised approach. Zhou et al. [156] study temporal transformations of a single object state in time-lapse videos. Wang et al. [139] propose to model state transformations in a high-level feature space with Siamese networks. Doughty et al. [36] leverage natural language and treat adverbs as modifiers for state transformations.

3) 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, action verbs, and nouns for the objects.

We annotate hand-object interactions corresponding to each narration within the selected 5 minute clips. We use the taxonomy from Section VII-B for semantic verb and noun labeling. The annotation pipeline consists of three sequential stages: critical frame labeling, pre-period labeling, and post-period labeling.

Given a narration, we create an 8 s video snippet centered at the narration time point and present it to the annotators. We ask the annotators to first read the narration and select a corresponding verb from the taxonomy. The annotators can then play the video back and forth to select three critical frames in time: PNR, PRE, and POST. We ask the annotators to start with the PNR frame that identifies the beginning of the state change. This frame is less ambiguous and helps provide the context for the interaction. We then ask the annotators to label a frame prior to the state change (PRE) and a frame after the completion of the state change (POST). Note that the PRE and POST frames are not uniquely defined. We let the annotators pick any, as long as the relevant objects are *fully* visible within the field of view of the camera.

Next, we label bounding boxes for the hands, tools, and objects, as well as the category names for the tools and objects. See Figs. 20 and 21. We do this in two steps. First we label the frames in the PRE period, starting at PNR and going backward to the PRE frame. The video frames are reversed and the annotators can play the video. We find that it is easier to start from the PNR frame since the hands and objects are clearly visible. To speed up hand box labeling, we initialize the hand boxes with a pre-trained object detector and ask the annotators to correct these.

Finally, we ask the annotators to label spatial annotations and categories for the POST frame. As before, we first present the annotators with the PNR frame. Note that in this case the PNR frame is already labeled which helps identify the hands and objects to label in the post frame.

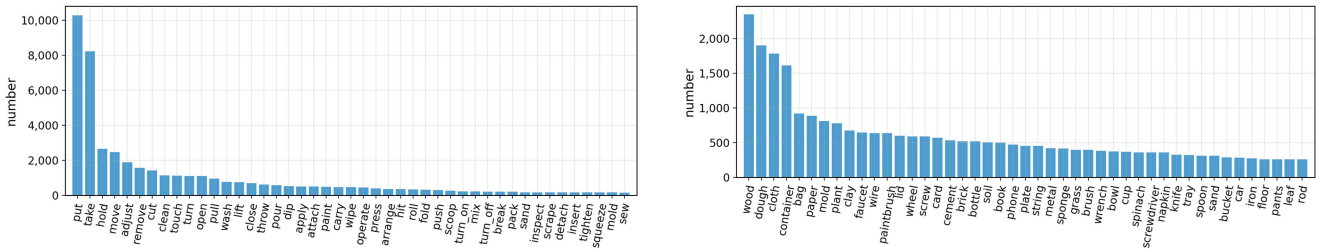


Fig. 18. *Labeled actions.* Distribution of verbs (left) and nouns (right) in annotated action instances. Top 45 verbs and nouns are shown for clarity. See Section VII-B for more details.

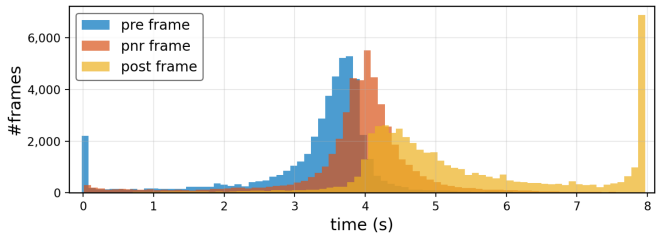


Fig. 19. *Critical frames.* Distribution of critical frame times. Shown relative to the 8 s hand-object interaction snippet.

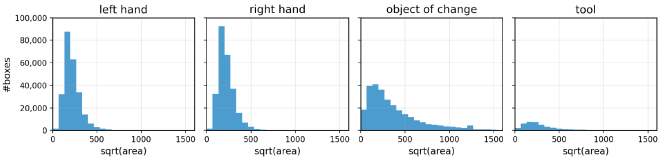


Fig. 20. *Hand and object sizes.* Distribution of bounding box sizes. Shown in terms of the square root of the box areas.

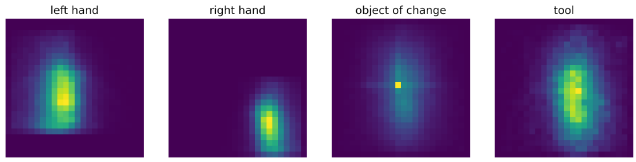


Fig. 21. *Hand and object locations.* Distribution of bounding box centers. Shown in normalized image coordinates.

To gain a better sense of the data used for the hands and object benchmark, we present analysis of our annotations. In Fig. 19 we show the temporal distribution of critical frames within the 8 s hand-object interaction snippets. First, we observe that the PNR frame distribution is centered around the middle of the snippets. Interestingly, this closely aligns with the narration point (4s mark). Next, we see that most of the pre and post frames come shortly before and after the PNR frame, respectively, highlighting the quick nature of these state changes, and thus the challenge in this benchmark. We also notice two additional modes for pre and post frames that come at the start and the end of the 8 s interval, respectively. These correspond to long repetitive actions that start before or continue past the video snippet (e.g., knitting).

We start with a large pool of videos annotated with high-level scenario labels (e.g., gardening, cooking, landscaping, etc.) and narrations. We assess each scenario on the scale of 0 to 3 based on how likely it is to contain hand-object interactions (e.g., 0 for "watching tv", 3 for "carpentery", etc.). We then sample data to annotate following the resulting scenario distribution. Given a scenario and a target number of hours, we sample clips randomly in a hierarchical fashion: we first sample a participant, then a video, and finally a 5 minute clip from the video. If the video is shorter than 5min we take the whole video. For each scenario, we balance the data across universities to maximize geographic diversity. The resulting scenario and university distributions are shown in Fig. 22. In total, our dataset has 120 hours representing 53 scenarios, 7 universities, and 406 participants.

4) Evaluation Metrics and Baselines: Point-of-no-return (PNR) temporal localization is evaluated using absolute temporal error measured in seconds. State change object detection is evaluated by average precision (AP). Object state change classification is evaluated by classification accuracy. The training, validation and test sets consists of approximately 19K/33K, 13K/22K, and 13K/22K, state changes / hand bounding box annotations, respectively.

The performance of several baseline methods for Temporal Localization and Object State Change Classification are given in Table V (top) and (bottom). The winner of the ECCV 2022

TABLE V
PERFORMANCE ON HAND-OBJECT TASKS

PNR Temporal Localization	Val (seconds)	Test (seconds)
Always Center Frame	1.032	1.056
BMN	0.780	0.805
I3D ResNet-50	0.739	0.755
Bi-directional LSTM	0.790	0.759
SlowFast + Perceiver	0.804	0.828
ECCV 2022 [153] (CSN, VideoMAE)	0.502	**0.516**

State Change Object Detection	Test (AP)	Test (AP-50)	Test (AP-75)
Faster-RCNN (ResNet-101)	13.4	25.6	12.5
DETR (ResNet-50)	15.5	32.8	13.0
CenterNet (DLA-34)	6.4	11.7	6.1
100DOH Model (ResNet-101)	10.7	20.6	10.1
ECCV 2022 [19] (Swin-L, DINO)	**37.19**	**55.97**	**38.44**

Object State Change Classification	Val (Accuracy)	Test (Accuracy)
Always Positive	48.1	47.7
Bi-directional LSTM	65.3	63.8
I3D (ResNet-50)	68.7	67.6
ECCV 2022 [153](CSN, VideoMAE)	77.2	**79.6**

Baseline method performance and Ego4D ECCV 2022 challenge winner performance for each tasks.

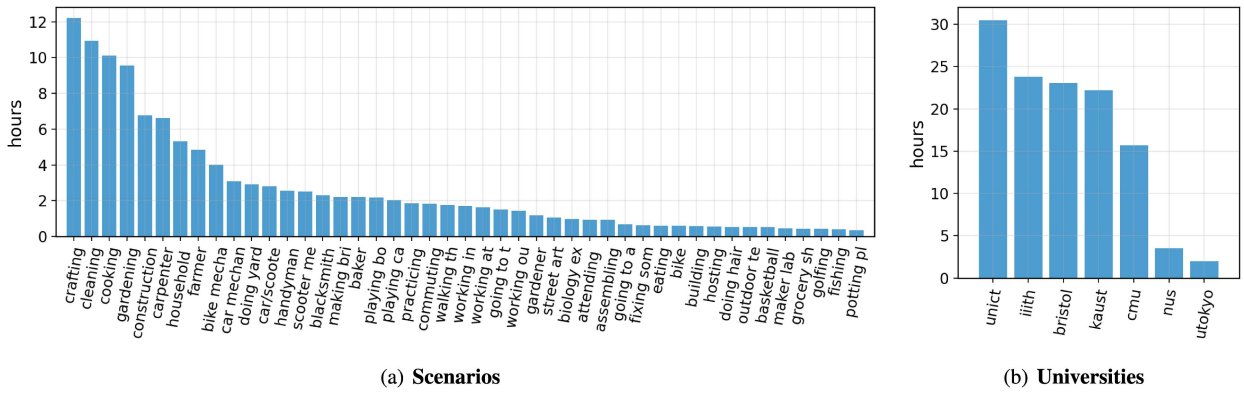


Fig. 22. *Number of hours.* We show the distribution of the number of hours of video in the annotated hands-object benchmark across scenarios (left) and universities (right).

challenge fused the result of a Channel- Separated Convolutional Network (CSN) and Video Masked Autoencoder (VideoMAE) detailed in [153]. They achieved top performance of 0.516 seconds error for the PNR temporal localization task and 79.6% accuracy for the object state change classification task. Key insights were the use of dense frame sampling and making use of the center-bias of the dataset.

Among the implemented baseline models, in general there are one or two types of output network heads: a classification head for the video clip used for state change classification, or a per-frame classification head for temporal localization. One can choose to train two models separately, or use the same backbone model but two network output heads and train the joint model with a multi-task loss function. The following baseline methods includes both types of model designs.

We use I3D with ResNet-50 as backbone architecture of the model for both the Object State Change Classification and the PNR Temporal Localization tasks. The ResNet backbone is followed by two network output heads: a state change classification head and a PNR temporal localization head. The state change classification head is produced by global average pooling on the entire spatiotemporal feature tensor followed by a classification layer. The PNR temporal localization head is produced by per-frame average pooling followed by a classification layer. The overall training loss of the model is the combination of the loss of two heads which are both cross-entropy loss for classification.

We use a Boundary-Matching Network (BMN) [83] as a baseline for the PNR Temporal Localization task. BMN is a temporal segment detection method based on confidence prediction of dense temporal segment proposals. We view the start of the video as the start of the temporal segment and Point-of-no-return I_{pnr} as the end of the temporal segment, so we can convert the problem of localizing Point-of-no-return I_{pnr} to the problem of detecting the temporal segment. In our implementation, BMN uses ResNet as the backbone model. Furthermore, BMN is only used for the PNR temporal localization task.

We implement a baseline model whose architecture consists of SlowFast and Perceiver for both object state change classification and PNR temporal localization. SlowFast acts as the video deep feature extractor. The features are then passed to a Perceiver model. Similar to the previous BMN model, the SlowFast + Perciever model is only trained for temporal localization. The training loss is the cross-entropy loss for per-frame classification.

We implement a Bi-directional LSTM model for both the object state change classification and PNR temporal localization. We first pass individual frames to a ResNet model to extract deep features. The sequence of per-frame features is then passed to the Bi-directional LSTM as input, with the output sent to both the per-frame classification head and the whole-sequence classification head. The overall training loss is the combination of the loss of two heads which are both cross-entropy loss for classification.

While we expect that new methods developed for the tasks of state change object detection will utilize all three input frames (PRE, PNR, POST), in this initial stage of the benchmark, we only evaluate single-frame detection baselines, where only the PNR frame I_{pnr} is used as input. We present the implementation of several baseline methods for the state change object detection task. In general, the baseline models for the task can be categorized into two types: (1) directly detecting the bounding box of the state change object including Faster-RCNN, CenterNet, and DETR, and (2) detecting hand bounding boxes first then predict state change object bounding boxes given the hands such as the 100DOH model [128]. Specifically, the baseline methods are the following:

Faster-RCNN is a two-stage anchor-based 2D object detector on a single RGB image. In its classification head, the state change object is the only positive category. We train Faster-RCNN on our benchmark and use it to directly detect the bounding boxes of state change objects in PNR frames.

CenterNet is another object detection method on a single RGB image. It estimates object keypoints to find object center points and regresses all other object properties, such as size, 3D location, and orientation. We train CenterNet to directly detect the bounding boxes of state change objects.

DETR is an object detection model on a single RGB image based on Transformer. It views object detection as a direct

set prediction problem and uses a transformer encoder-decoder architecture to produce a set of object predictions including bounding box information as well as other information such as category. We train DETR to directly detect the bounding boxes of state change objects.

100DOH Model [128] first detects the bounding boxes of the human hand and objects as well as the relational vectors that links from each hand bounding box center to an object bounding box center. The final prediction of the objects are decided as the object predictions that satisfies the both the predictions of hand and relational vectors. We used the 100DOH model pre-trained on 100DOH dataset to first detect hand bounding boxes and then predict state change object bounding boxes given the hands.

The results of single-frame State Change Object Detection are illustrated in Table V (middle). All baselines struggle in detecting the State Change Objects with only one frame as input with an AP of 6-15%. There are several challenges. First, the bounding box sizes of state change objects have large variance. For example, the size of state change objects can be as large as half of image in the action of "painting the wall" and as small as a few pixels in the action of "igniting the match." Second, when only using one frame as input, the detection models did not consider the change of object appearance across different frames. The ECCV 2022 challenge winner [19] followed baseline methods using only the PNR frame and utilized a Swin transformer and DETR-based detection head, along with model pre-training, and improved performance to 37.19 AP. Moving forward, we hope the researchers will investigate using models that take multiple frames as input and develop frameworks that incorporate tracking or association.

5) Discussion: The hands and object benchmark has proposed three new tasks that can be used to evaluate the ability of a model to understand how hands induce state changes for objects. The ECCV 2022 challenge winners have shown the importance of model pre-training and using more powerful models to characterize object state changes. At the same, they have also helped to identify and emphasize the next phase of research for object state change understanding. While models can localize PNRs in time, it is not clear if these model are actually encoding the transformation of an object over time. Models can be optimized to detect dominant objects from the egocentric view given a PNR frame, but it is not clear if they understand how the object is undergoing a state change (i.e., current methods do not make use of PRE or POST images). Top performing object state change classifier methods use sparse uniform sampling of video clips, and it is unknown if models can benefit from paying closer attention to key frames such as the PNR.

C. Audio-Visual Diarization

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.

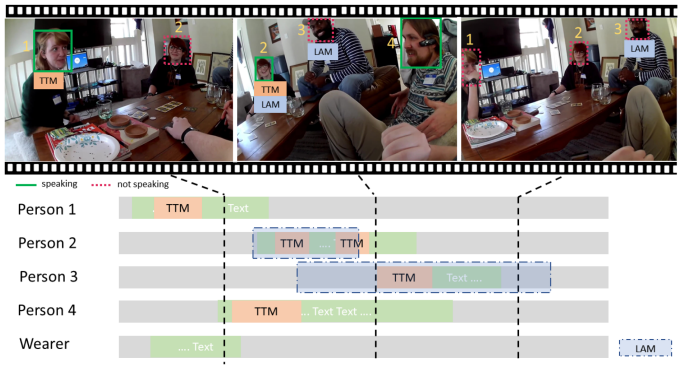


Fig. 23. Audio-Visual and Social benchmark annotations.

The goal of this Audio-visual diarization (AVD) benchmark is to help advance the state of the art in audio-visual understanding from the egocentric viewpoint. Specifically, from a conversational perspective, the benchmark aims to understand *who is talking when, and about what*. From a visual perspective, we are also interested in *where the speaker is located*. Given an egocentric video, the proposed tasks require extracting the spatial location of the speakers, their voice activity across the length of the video, and the content of their speech.

1) Task Definition and Annotations: This benchmark is composed of four tasks (see Fig. 23).

1) *Localization and tracking:* This task captures the spatial position of all the probable speakers in the scene, from the point of view of the camera wearer. The goal is to both localize and track such candidate speakers in the video frames. Unlike classical face detection benchmarks, this is challenging in the sense that the dynamics of the camera wearer's head (coming from natural conversations) leads to significant movement in a speaker's apparent spatial location.

As annotations, for each speaker present in a 5 minute clip a bounding box is provided. We first used a face detection and tracking model to estimate these bounding boxes, and then human annotators validated and corrected these machine-generated boxes to improve annotation quality. A bounding box is considered a valid human annotation if it captures 80% of the speaker's face. Sideways looking faces are also annotated. Note that speakers who are very far from the camera wearer (i.e., several meters away in the scene) and who do not come into conversational contact with the wearer are not annotated.

2) *Active speaker detection:* This task aims to detect the active speaker in the scene. It builds on top of the previous localization and tracking task to recognize each of the speakers whose face bounding boxes are detected. This task only takes into account speakers visible in the camera's field of view (FoV) plus the camera wearer.

As annotations, we provide an anonymous speaker label (e.g., speaker 1, 2 etc.) for each speaker visible in the clip. The camera wearer is assigned the label C. This is done by utilizing the face

bounding box tracks annotations and labeling each track one at a time. Hence, each face track gets assigned one unique label, and multiple tracks within a single clip may share the same label (corresponding to the same speaker). However, the labels are clip-specific, i.e., a speaker who may be present across multiple clips does not get assigned a shared unique label across the clips. Aside from the camera wearer, speakers who are never in the visual FoV are not assigned a label.

3) *Diarization:* Given the set of speakers and their spatial localization, the diarization task aims to capture the voice activity of the speakers, answering the question "who spoke when". While speech from speakers that overlap with each other is one of the biggest issues to solve in this task, the egocentric perspective adds more complexity in terms of head motions and other dynamics associated with natural conversations.

For every active speaker label (where the annotations are from the previous active speaker detection task), a human annotator marks the start and end time of that person speaking. We account for overlapping speech segments where multiple speakers talk over each other, but we ignore speech not relevant to the conversation such as background speech from a TV or speech further away from the camera wearer. Note that speech segments from the camera wearer are also annotated. The annotators rely both on the audio and the visual stream for creating these labels.

4) *Transcription:* The final task of AVD is to transcribe the speech of each speaker. Similar to the diarization task, some of the challenges associated with the transcription task include overlapping speech and environmental noise. In addition, the camera wearer's head movement results in a significant change of the audio volume of the speech recorded from others.

The transcriptions are obtained in multiple passes to ensure overlapping speech segments are accounted for correctly. In the first pass, human annotations based on voice segments are merged with automatic annotations for regions with low volume. In a subsequent pass, human annotators correct and assign segments of transcriptions to the corresponding voice activity segments per speaker while also annotating overlapping speech. Annotators are provided both the audio and video for annotation. Besides spoken words, the occurrence of other artifacts such as unintelligible speech or incomplete words are also annotated. The final annotations consist of a sequence of segments labeled with begin time, end time, transcript and speaker ID within the clip. Note that the time segments associated with the transcripts are not the same as the ones used in diarization because we separately annotate the overlapping regions here to reduce transcription errors and account for speakers talking at a low volume. This allows us to also distinguish voice activity from speech activity.

From across the 3,670 hours of video in Ego4D, approximately 764 hours of data contains conversational content, and are directly relevant for the AVD and Social benchmarks. As discussed above, capture for the audio-visual social settings used in these benchmarks entailed closed environments and informed consent. From this set, a randomly chosen subset of 572 clips (each 5 minutes long) are annotated for the v1 release. Of these 572 clips, 389 clips (32.4 hours) are marked for training, 50 clips (4.2 hours) for validation, and the remaining 133 clips (11.1 hours) is the testing set, totaling 47.7 hours of annotated data.

2) Evaluation Metrics: We use standardized object tracking (MOT) and Identity (ID) metrics [9], [10] to evaluate speaker localization and tracking. Multiple object tracking accuracy (MOTA) is a combined metric of false alarms, false positives, and identity switches, based on matching the tracking results with the ground truth at the frame level. The ID precision (IDP), ID Recall (IDR), and ID F1 score (IDF1) are based on the tracking result to ground truth matching at the trajectory level. For evaluating active speaker detection, we use the object detection mAP: in a video frame, if the intersection over union (IoU) between a detected face bounding box and the ground truth face bounding box exceeds a predefined threshold, i.e. 0.5, we have a positive face detection. Diarization error rate (DER) is the *de facto* evaluation metric for speaker diarization [6], and it is well studied in the audio and speech processing community. DER measures the fraction of total time (in a given clip) that is not attributed correctly to a speaker or to non-speech: DER (%) $= (E_{miss} + E_{fa} + E_{spk}) \times 100$, where E_{miss} denotes the fraction of time that has been predicted to be non-speech while that segment is attributed to a speaker in the reference, E_{fa} denotes the fraction of time that has been predicted to be associated with a speaker, but is actually labeled as non-speech in the reference, and E_{spk} denotes the fraction of time where speech is associated with the wrong speaker. All errors are computed as a fraction of the total amount of speech. Lastly, we use the Word Error Rate (WER), a standard ASR metric [68]. First, the minimum edit or Levenshtein distance is computed between the reference and hypothesized transcription. WER then measures the ratio of the number of word substitutions (S), deletions (D) and insertions (I), i.e. the total number of edits necessary to convert the hypothesized transcription into the reference relative to the total number of words (N_w) in the reference:

$$\text{WER}\ (\%) = \frac{S + D + I}{N_w} \times 100. \tag{1}$$

3) Relation to Existing Tasks and Data: VoxCeleb 1 and 2 [24], [102] have recordings of more than 6K speakers spanning a wide range of different ethnicities, accents, professions, and ages. The data is non-egocentric and is annotated for active speaker face bounding boxes, face tracks, and anonymous person IDs. VoxConverse [23] is a related audio-visual diarization dataset consisting of over 50 hours of multi-speaker clips of human speech, extracted from YouTube videos. Similar to VoxCeleb, this data is also non-egocentric. The AVA spoken activity datasets are AVA speech and AVA active speaker [17], [124]. AVA speech is a densely annotated audio-based speech activity collection of AVA 1.0 third-person videos, and explicitly labels three background noise conditions, resulting in approximately 46,000 labeled segments spanning 45 hours of data. AVA active speaker associates speaking activity with a visible face, resulting in 3.65 million frames labeled across approximately 39,000 face tracks. The closest egocentric dataset for audio-visual diarization is AVDIAR [54]. It consists of 23 staged sequences, with each sequence duration ranging from ten seconds to three

minutes (a total of 27 minutes of video). EPIC-Kitchens has 100 hours of video with audio in non-scripted recordings in native environments, but does not have speech- or speaker-related annotations. EASYCOM [34] is an egocentric dataset with five hours of multi-channel data for conversational content with $3-5$ participants in a closed room setting.

Recent work explores audio in computer vision tasks [157] for action classification [66], [142], object categorization [73], [147], source localization and tracking [7], [127], [136] and embodied navigation [18]. On audio-visual detection and tracking, recent work explores ways to localize sounds in a given video frame [7], [127], [136] and infer spatialized sound from video [51], [99]. Capturing and processing multi-channel audio is being studied in audio and microphone array signal processing communities, specifically from a user's perspective to understand a given scene [60], [107]. Recent work shows that audio disambiguates certain visually ambiguous actions [66], [142].

Audio-visual speech recognition has received a lot of attention in the last decade, with multiple studies suggesting that automatic speech recognition (ASR) can benefit from visuals of the scene, or other non-acoustic information [3], [62]. In addition, audio-visual cross-modal learning benefits the cocktail party problem, which requires recognizing what one person is saying when others are speaking at the same time [4], [38], [50], [52], [53], [54], [109], [152].

4) Baseline Models: Recall that the four-part tasks in this benchmark are tied to each other, in the sense that representations learned from one task may be relevant for the others. To that end, we propose a baseline that addresses each task in a sequential fashion. The framework includes the following steps: (1) We first detect people's heads and do short-term tracking in the video. The short-term tracker follows each detected head by expanding a set of trajectories based on their positions, sizes, and the appearance of the person. The trajectories may end when occlusion happens or when the tracked person goes out of the field of view. New trajectories can also be added to the trajectory set. (2) The short-term tracker's trajectory for each person is often fragmented into multiple parts. Hence, we then optimize the grouping of the tracklets from above so that the trajectories of each person can be linked together. We formulate the problem as a constrained combinatorial optimization problem. Integer programming can be used to solve the problem directly but it has exponential complexity. For efficiency, we develop a greedy approach which is much faster and still gives strong results. (3) We then classify each person/head in each video frame as an active speaker or not. Based on the classification result and the corresponding detected long-term trajectories, we further associate the audio/speech to each person in the video. We use this preliminary list of audio feature embeddings to further extract and match un-associated audio segments to speaker labels. (4) We then use two methods to detect the camera wearer's voice activity. The first method uses the high energy audio segment in the clip (under the assumption that their voice has natural amplification compared to the remaining speakers). The second method is a deep classifier that predicts whether the wearer is speaking. (5) Lastly, we apply ASR to the speech regions based on the ground truth segmentation and evaluate the WER across all segments. We used pre-trained Gigaspeech model provided in the ESPNet model zoo [2]. We leave jointly modeling the time segments and transcriptions as a future challenge.

TABLE VI
BASELINE PERFORMANCE ON TESTING SET FOR THE AVD TASKS, AND THE WINNING TEAMS/MODELS FROM ECCV 2022 CHALLENGES

Model	Metric	Test
Localization/Tracking		
Short-term tracklets + Long-term concatenate (ref, Sec 6.3.4)	MOTA IDF1	71.94 80.07
Active Sp. Detection		
RegCls + max-filtering + sVAD (pretrain)	mAP	33.72
RegCls + max-filtering + sVAD (No-pretrain)		34.35
TalkNet + sVAD (pretrain)		34.56
TalkNet + sVAD (No-pretrain)		49.66
Diarization		
RegionCls (No-pretrain)	DER	80.52
RegionCls (No-pretrain) + sVAD		80.17
TalkNet (No-pretrain)		73.14
TalkNet (No-pretrain) + sVAD		73.32
Spell-based @ Eccv22		65.9
kVAD (Audio-Only)		65.28
pyannote @ ECCV22 (Audio-Only)	50.7	
Transcription		
Naive baseline	WER	112.10
AVATAR @ ECCV22		68.40

Active speaker detection is an important component that influences the diarization as well as transcription tasks. Hence, to ensure the baselines we build are strong enough with regard to exploring the inter-dependancy of tasks, we utilize two approaches. Our first approach called RegionCls is based on the classification of mouth regions. It first computes the 3D head orientation using a regression network. If the face is facing away from the camera, we ignore the image and the active speaker detection result is set to null. For faces looking at the camera, we regresses the facial key points using the image within the person's head bounding box. We use the mouth key points to crop out the mouth image. The cropped mouth image is then sent to a classification network to classify whether the speaker is talking or not.

In addition, we use TalkNet [135] which is an end-to-end pipeline that takes the cropped face video and corresponding audio as input, and decides if the person is speaking in each video frame. It consists of a feature representation frontend and a speaker detection backend classifier. The frontend contains a frame-based audio and video temporal encoders. The backend classifier consists of an inter-modality cross-attention mechanism to dynamically align audio and visual content via self-attention. Note that any video-only approach for sound activity can be combined with a voice activity detector to remove false alarms. Here we use such an algorithm from [130] – referred to as sVAD to improve the active speaker detection results.

Table VI summarizes the resulting performance metrics for the tasks. In addition to the baseline framework discussed earlier, Table VI also shows the winning models from ECCV 2022 challenges on audio-visual diarization. This includes SPELL based active speaker detection to improve audio-visual diarization [96], neural speaker embeddings driven audio-only

diarization [14], and audio-visual ASR based transcription prediction model [49].

5) Discussion: Although AV diarization presents a task suite composed of reasonably well understood tasks from the vision, speech and audio communities, our baseline results clearly suggest that efficient speaker localization, tracking, diarization and transcription is a rather complex problem in the egocentric perspective and with in-the-wild data. This is specifically evident from the performance of the joint audio and video driven diarization and transcription baselines (with DER of $> 80\%$ and WER of $> 60\%$). Overlapping speech makes both these tasks particularly difficult to annotate as well as evaluate any proposed models. Performing some audio-visual source separation prior to these tasks may improve the efficacy; nevertheless sensitivity to changes and difference in speech amplitudes of overlapping speakers would still be challenging to address. In addition, the relationship between robust localization and tracking with multi-speaker diarization is not studied, and this is also not well understood in the literature. We expect this to be a challenging problem.

Novel cross-modal learning approaches that jointly model audio and visual modalities while accounting for such attributes (overlapping speakers, interruptions, noise in the wild etc.) are needed to further improve these performances. This was shown already by the winning models of ECCV 2022 challenges wherein robust audio-visual representations drove the bulk of the gains in performance. We also observed that subjective attributes in conversations, like speaker accents, as well as changes in vocabulary usage based on cultural differences influence both the content of the speech and the clarity with which it can be captured in human annotations. The camera wearer's head motion adds significant blur to speakers' faces. To account for such aspects we performed quality checks on human annotations. We expect novel unsupervised and self-supervised learning will help further address such subjective attributes.

D. Social Interactions

An egocentric video provides a unique lens for studying social interactions because it captures utterances and nonverbal cues [69] 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 [120], and could support novel prosthetic technologies for the hearing-impaired. We next present our Social Interaction benchmark.

1) 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.

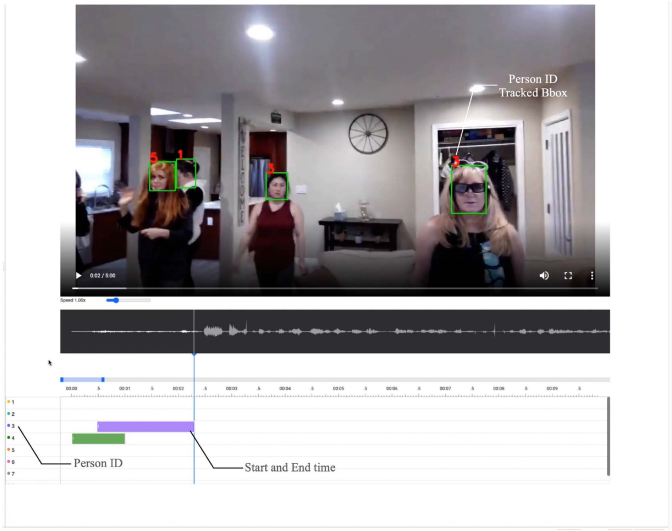
Formally, LAM and TTM are defined as follows: (1) LAM: $y = f(\mathbf{I}, \mathbf{B})$; (2) TTM: $y = f(\mathbf{I}, \mathbf{A}, \mathbf{B})$ where $\mathbf{I} = \{I_t\}_{-T_1}^{T_2}$, $\mathbf{A} = \{A_t\}_{-T_1}^{T_2}$, and $\mathbf{B} = \{B_t\}_{-T_1}^{T_2}$ are time-synchronized past sequences of video, audio, and bounding boxes, respectively, where T_1 and T_2 are the length of the past and future time horizon, respectively, and $t = 0$ is the center frame. The bounding box indicates the target person to classify. y is a binary classification label defined by:

$$y = \begin{cases} 1 & \text{if target looks/talks at camera wearer} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

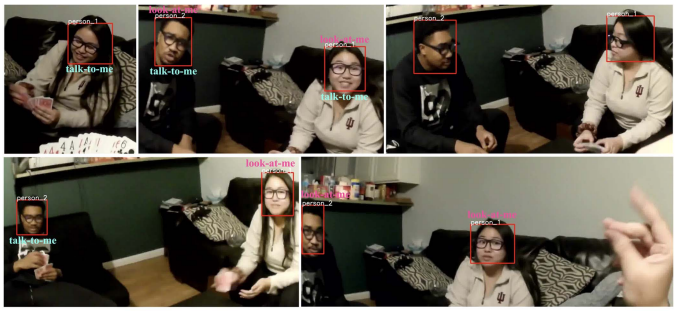
The LAM and TTM tasks are defined as a frame-level prediction y, which stands in contrast to audio analysis tasks where labels are often assigned at the level of audio frames or segments. A desired model must be able to make a consolidated decision based on the video and audio cues over the time course of an utterance. For example, if the speaker turns their head to the side momentarily while speaking to the camera-wearer, then a frame where the speaker is looking away would have $y_{\text{LAM}} = 0$ while $y_{\text{TTM}} = 1$. Fig. 24 gives some frame level visualization of annotations that illustrate the task definitions.

2) Annotations: The Ego4D Social data collection process was designed to achieve: 1) naturalistic interactions, 2) multimodal capture, and 3) diverse participants and environments. Participants consisted of friends and family groups and data was captured in residences and local neighborhoods, ensuring naturalistic interactions. Capture hardware varied across sites but included wearable cameras, wearable eye trackers at Georgia Tech and Indiana University, binaural recording systems, and smart watches at Georgia Tech. Protocols included highly-structured settings, where participants were asked to play games over a period of a few hours in a residence, and unstructured settings where participants captured social interactions in daily life over a period a week or more. Sample social interaction contexts included playing board and card games, preparing meals, and going on walks. The bulk of the data collection took place during the COVID-19 pandemic, and the resulting study protocols were designed to safeguard participants against additional risk. The Social data consists of data collected at five sites: Atlanta, Bloomington, Redmond, Twin Cities, and Singapore.

Social annotations build on those from AV diarization (Section VI-C). 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 Fig. 23.



(a) Annotation tool

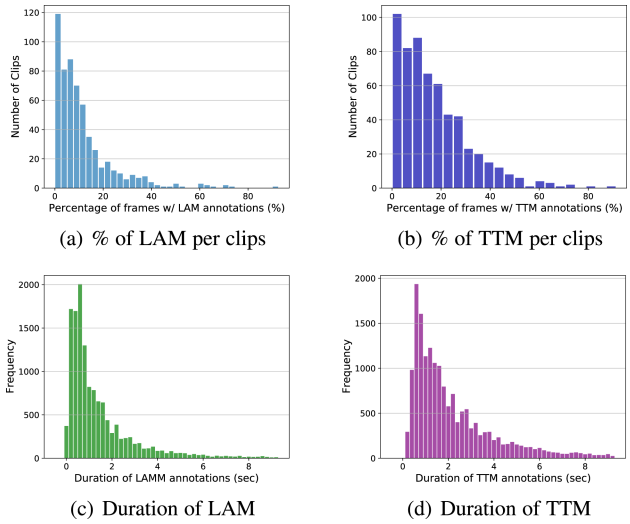


(b) Visualization of annotations.

Fig. 24. (Top) The GUI of the annotation tool; (Bottom) Visualization of example annotations. Note that LAM (denoted by magenta text) and TTM (denoted by cyan text) may not necessarily occur together as shown in these examples.

Fig. 25 summarizes the statistics of LAM and TTM annotations. We compute the percentage of the frames with LAM or TTM annotations in each clip and show the histograms in Fig. 25(a) and (b), respectively. In many clips, these events happen rarely (10 % or lower), and the frames with LAM annotations are less frequent than TTM cases. We also list the duration of each LAM or TTM annotation in Fig. 25(c) and (d), in order to illustrate the significant variations in length. The most frequent case is short-duration LAM or TTM behaviors, lasting 1 or 2 seconds. The same data split described in Section VI-C1 was used for model training: 389 clips (32.4 hours) were held out for training, 50 clips (4.2 hours) for validation, and 133 clips (11.1 hours) for testing.

3) Relation to Existing Work: Compared to [43], 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 ego-video [20], [98], but addresses more diverse and challenging scenarios. Mutual gaze estimation [35], [90], [91], [92], [110], [112] and gaze following [21], [42], [67], [119] are also relevant. The TTM task is related to audio-visual speaker detection [5], [124] and meeting understanding [12], [76], [94].



(a) % of LAM per clips (b) % of TTM per clips

(c) Duration of LAM (d) Duration of TTM

Fig. 25. Social task annotation statistics. (a) Histogram showing the number of clips vs. the percentage of frames with looking-at-me annotations. (b) Histogram showing the number of clips vs. the percentage of frames with talk-to-me annotations in each clip. (c) Histogram showing the duration of looking-at-me (LAM) annotations. (d) Histogram showing the duration of talking-to-me (TTM) annotations.

4) Evaluation Metrics and Baselines: We use mean average precision (mAP) and Top-1 accuracy to quantify the classification performance for both tasks. We calculate these metrics at the frame level.

LAM Our baseline model for LAM is a video-based model using ResNet-18 and Bidirectional LSTM. Our model uses the cropped face regions in video as input in order to focus on cues about the head pose and social attention visible in the face. The architecture of our baseline is similar to the Gaze360 [67]. As illustrated in Fig. 26(a), we input seven consecutive frames ($T_1 = 3$ and $T_2 = 3$) from one face tracklet, and each image is resized to 224×224. Each frame is then processed by the ResNet-18 backbone independently to generate 256 dimensional face features. The feature sequence is encoded by a Bidirectional LSTM, which has two recurrent layers with dimensionality 256. The output is fed into a classification head to predict the binary LAM result for the center frame at the t-th timestamp. The LAM task has a class imbalance issue, and we use weighted cross-entropy loss. Since the architecture is similar to Gaze360, we have two options for the initialization: first, initializing the backbone from a pretrained Gaze360 model; second, initializing the model randomly and training from scratch on Ego4D. During training, we sample center frames with a stride of 3. The network is optimized by Adam with a learning rate of 5×10^{-4}.

Table VII shows the LAM results. Our baseline model achieves a mAP of 66.07% on the test split when initialized randomly, and the performance is higher at 72.11% when initialized from Gaze360. These findings highlight the close relationship between the LAM task and gaze estimation. The random guess model achieves about 8% accuracy because the negative samples account for 92% of the test split and the model always predicts looking at me.

TTM The baseline model for TTM digests multi-modal inputs: each audio segment is paired with an associated face crop.

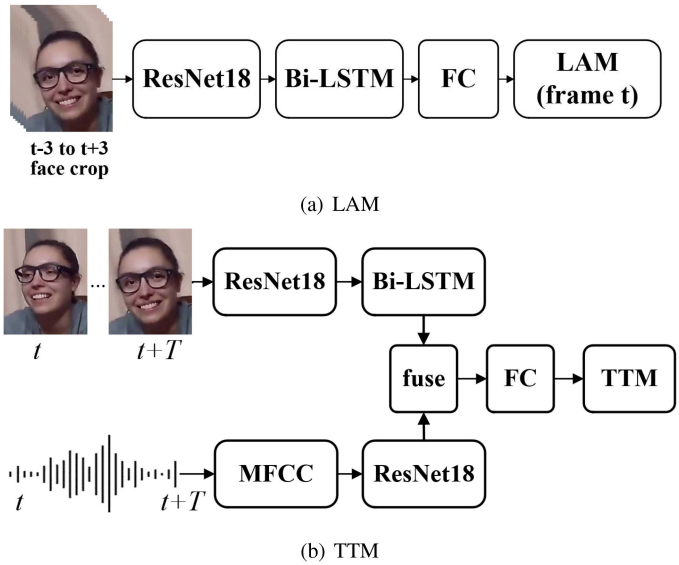


Fig. 26. *Baseline model architectures.* (a) LAM model uses a ResNet-18 as a backbone to extract the feature of each frame. A Bidirectional-LSTM then takes the sequence and encode the features into one embedding. We pass the embedding to FC layer that predicts the LAM result. (b) TTM model has two encoders. The video encoder is the same as LAM. The audio encoder extracts the MFCC frequency map of the audio segment and the feature is fed into a ResNet-18 network. The visual and audio embeddings are concatenated and passed through the FC layer to predict the target of this utterance.

TABLE VII
THE RESULTS OF THE LAM AND TTM TASKS

LAM Results

	val		test	
	Acc	mAP	Acc	mAP
Random Guess	8.57	51.19	7.98	50.96
Ego4D Baseline (Gaze360)	91.78	79.90	86.45	72.11
Ego4D Baseline (Random)	87.97	78.07	75.38	66.07
Winner (PKU-WICT-MIPL)			93.62	78.07

TTM Results

	val		test	
	Acc	mAP	Acc	mAP
Random Guess	32.44	53.82	47.41	50.16
Ego4D Baseline	64.31	56.50	49.75	55.06
Winner (icego)			55.93	57.52

For LAM, we report the baseline initialized from Gaze360 [67] (2nd row) and randomly (3rd row). The baseline model for TTM is initialized randomly. We also include the test set performance of the winning entry for each task from the Ego4D challenge held at ECCV 2022.

Since the audio segments vary substantially in duration, we break the long utterances into short segments whose maximum duration is limited to 1.5s. If the segment is shorter than 0.15s, we skip it in the training stage. The associated faces are also resized to 224×224, and the video encoder is the same as LAM. However, sometimes the speakers leave the field of view or become invisible due to the rapid motion. In this case, we pad the face sequences with blank images. The MFCC feature is extracted every 10ms with a 25ms window length. The feature is then fed into the audio backbone, a ResNet-18 designed for audio tasks [22]. Following the encoders, we concatenate the audio and visual embeddings and pass them to the final classification head to get the TTM result for the visible faces associated with the segment. To train the model in parallel, we first sort the short segments based on the length and group the segments into a batch if they have the same duration. The batch size is restricted by the GPU memory; we use a batch size of 400. The model is also optimized using Adam with a learning rate of 5×10^{-4}.

Table VII shows the TTM results. TTM is more challenging than LAM. We can see that our baseline model only increases the mAP by 9.77% on the test split in comparison to the random guess model.

5) Discussion: While the benchmark tasks of detecting when attention and speaking behaviors are directed towards the first-person are closely related to existing analysis tasks, it is clear from the baseline performance that there is substantial room for improvement.

The TTM task is particularly challenging because it requires analysis of the audio content to understand the target audience of an utterance, as well as the fusion of audio and video cues. The most complete solution to this problem will require an understanding of the semantics of the utterance in the context of an evolving conversational interaction. Future work on this task might involve more sophisticated language modeling and possibly hierarchical analysis approaches that allow the integration of cues at multiple levels, e.g. at the dialog level to understand who is participating in a conversational exchange, at the utterance level to access semantics, and at the audio level to exploit prosodic and other cues. The LAM task presents additional challenges such as the need to deal with motion blur and fast head movements, and may also benefit from a more explicit modeling of head movement and the patterns of gaze behavior that arise in conversational interaction.

The core tasks of LAM and TTM define a starting point for analyzing multi-modal egocentric data and inferring social interactions. We now describe two groups of potential future tasks that could be supported by Ego4D data.

Egocentric attention prediction: Prior work [78], [79] has demonstrated the feasibility of predicting where the camera-wearer is looking (i.e. their egocentric attention) using only egocentric video captured from a head-worn camera. This work leveraged the context of hand-eye coordination tasks, which require gaze to be coordinated with hand movements and objects. A subset of the Ego4D Social data includes gaze measurements produced by wearable eye trackers by Indiana University and Georgia Tech participants (e.g., Pupil Invisible), which will greatly expand the size of data for hand-eye coordination in the wild.

Social gaze prediction: The LAM task addresses the special case of social gaze: a person looks at the camera-wearer. It is possible to generalize the task by predicting the social gaze target for each of the visible faces in an egocentric video. including non-social gaze targets (e.g. looking at an object), looking at people who are not wearing an egocentric camera (with the result that ground truth annotations are not available), and looking at unknown targets not captured in any of the egocentric videos. The Ego4D Social data includes synchronized videos from multiple social members, which would expanding the annotation by matching the person ID with the camera-wearers. Note that since the video recorders are not genlocked, the identification of

corresponding frames will only be approximate. However, since gaze behaviors persist over multiple frames we do not believe this will be an issue.

A key issue in defining this future task is the determination of the participant set. For a 2D version, the participants are those who are visible in frame t. This is a social version of the video-based gaze following task [21], where the goal is to predict whether each target participant is looking at any of the other participants who are visible in the frame. A more challenging 3D version of the task would use all of the participants who are present in the social scene at the time of frame t. This task requires the ability to predict which participant the target person is looking at in the case where that participant is not visible in frame t. This can in principle be accomplished by maintaining a birds-eye view layout map of the social scene that captures the approximate spatial relationships between the participants. Such a layout map could be used in conjunction with an approach like Gaze360 [67]. Note that this task could potentially benefit from taking recorded binaural audio as an additional input, as the ability to localize sound sources could provide additional cues for determining the locations of gaze targets which are not visible in the video.

Utterance target prediction: The TTM task can be generalized to the full set of participants in the same way that LAM can be extended above. The input space is the same as TTM and the output space is similar to social gaze prediction, where we would infer whom the target participant is talking to, if anyone, or whether they are talking to someone who is not wearing an egocentric camera (and therefore ground truth cannot be determined). In contrast to social gaze, utterance target prediction requires the identification of all of the target recipients of an utterance. In fact, our TTM annotation already supports this task, as it differentiates the case where the utterance is directed to multiple participants including the camera wearer. This additional label is ignored in the design of the simpler TTM task.

Transcript-based variants: For all of the previously-defined social tasks it is possible to define a variant which utilizes a transcript of the audio file as an additional input modality. For example, the TTM-T task is the variant of TTM with the prediction defined as $y^p = f(\mathbf{I}, \mathbf{A}, \mathbf{T}, \mathbf{B})$, where $\mathbf{T}$ the transcript (time-stamped sequence of words) obtained from $\mathbf{A}$. This can potentially simplify the use of dialog cues to identify the intended targets for utterances and social gaze.

E. Forecasting

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, 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 world.

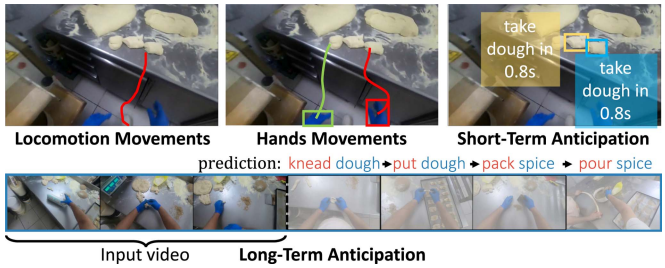


Fig. 27. The Forecasting benchmark aims to predict future locomotion, movement of hands, next object interactions, and sequences of future actions.

1) Task Definition: The Forecasting benchmark includes four tasks:

- *Locomotion prediction:* given a previous few seconds of an input video clip, predict a set of possible future ground plane trajectories of the camera wearer.
- *Hand movement prediction:* given an input video clip, predict the hand positions of the camera wearer in future frames.[12]
- *Short-term object interaction anticipation:* given an input video clip, 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.
- *Long-term action anticipation:* given an input video clip, predict the camera wearer's future sequence of actions performed by the camera wearer in the future.

Fig. 27 illustrates the aforementioned tasks.

2) Relation to Existing Tasks: Predicting future events has increasing interest [123]. Previous work considers future localization [111], action anticipation [48], [55], next active object prediction [11], [47], future event prediction [105], and future frame prediction [86]. Whereas past work relies on different benchmarks and task definitions, we propose a unified benchmark to assess progress in the field.

3) 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 (see Section VII-B). 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 $1.5s$. The hands have been also annotated in the same frames. We obtain ground truth ego-trajectories of the camera wearer using structure from motion.

4) Evaluation Metrics: Locomotion prediction: We measure the accuracy of the prediction using two metrics:

1) K best mean trajectory error (K-MTE). We measure K best trajectory error:

$$\text{K} - \text{MTE} = \underset{\{\mathcal{X}_k\}_{k=1}^{K}}{\operatorname{argmin}} \frac{1}{\sum_t v_t} \sum_t v_t \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|, \quad (3)$$

[12]Key frames include contact frame, pre-condition frame and three frames preceding the pre-condition frame by 0.5 s, 1 s and 1.5 s

$\mathbf{x}_t \in \mathbb{R}^2$ is the predicted location at time t, $\widehat{\mathbf{x}}_t$ is the ground truth location, and v_t is the visibility. The visibility indicates the availability of the ground truth trajectory, i.e., due to severe egocentric videos, the ground truth trajectories may include missing data. $v_t = 0$ indicates missing data at time t.

2) Probability of correct trajectory (PCT). We measure the success rate of the correct trajectory retrieval:

$$\mathrm{PCT}\epsilon = \frac{1}{K}\delta\left(\frac{1}{\sum_t v_t}\sum_t v_t\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\| < \epsilon\right), \quad (4)$$

where $\delta(\cdot)$ is one if the statement is true and zero otherwise. ϵ is the trajectory error tolerance, i.e., if the trajectory error is smaller than the error tolerance, it is considered as a correct trajectory prediction. $\mathrm{PCT}\epsilon$ measures how many trajectories among K retrieved trajectories are close to the ground truth trajectory.

Hand movement prediction: As for the future hands movements prediction, we only consider the key frame prediction, and therefore adopt Mean Key Frame Displacement Error Contact (M.Disp.) Key Frame Displacement Error as evaluation metrics (C.Disp.):

- Mean Key Frame Displacement Error (M.Disp.):

$$D_m = \frac{1}{n}\sum_{i \in H_t} \|h_i - \hat{h}_i\| \quad (5)$$

H_t refers to the set of visible hand positions of key frames, and n is the length of set H_t. h_i denotes the predicted hand position in the image coordinate, while $\hat{h}^i$ denotes the ground truth hand positions.

- Contact Key Frame Displacement Error (C.Disp.):

$$D_c = \|h_c - \hat{h}_c\| \quad (6)$$

h_c refers to the hand positions at Contact frame.

Note that all reports are reported on downsampled video frames with height of 256 and original aspect ratio.

Short-term object interaction anticipation: Methods will be evaluated at the timestamps in which next-active objects have been annotated, i.e.,

$$\left\{ t|t = t_s - l \cdot \alpha \right.$$
$$\forall t_s \in \{t_s^{(j)} | \exists \overline{h} : B_{\overline{h}}^{(j)} \neq \emptyset\}_j$$
$$\left. \forall l \in \{1, \ldots, m\} \right\} \quad (7)$$

where $\{t_s^{(j)} | \exists \overline{h} : B_{\overline{h}}^{(j)} \neq \emptyset\}_j$ is the set of all timestamps indicating the beginning of an interaction, for which at least one next active object has been annotated, m indicates the number of frames preceding the beginning of the interaction in which objects are annotated, whereas α is the temporal distance between the sampled frames.

Since detecting next active objects is a major part of the task, we base our evaluation measures on the standard Pascal VOC mean Average Precision (mAP). As in standard mAP, we first match each of the detected next active objects to ground truth annotations. A predicted and a ground truth bounding boxes are a possible match if their Intersection Over Union (IOU) value exceeds 0.5 and if some matching criteria are met. Predictions are matched to ground truth annotations belonging to the same evaluated example in a greedy fashion, prioritizing predictions with higher confidence scores and choosing matches corresponding to larger IOU values. A ground truth annotation can be matched at most with one predicted box. All matched predictions are counted as true positives, whereas all unmatched predictions are counted as false positives. Performance on the whole test set is summarized using the mean of the Average Precision values obtained for each class.

To account for the multi-modal nature of future predictions (i.e., more than one next active object can be likely), we "discount" the number of false positives obtained in a given example by the number of available ground truth annotations in that example multiplied by $K-1$, where K is a parameter of the evaluation measure. Specifically, if an example contains two ground truth annotation, we ignore the $(K-1)*2$ false positives with the highest scores. This effectively implements a "Top-K mean Average Precision" criterion which does not penalize methods for predicting up to $K-1$ possibly likely next active objects which are not annotated. Given a generic prediction $(\hat{b}_i, \hat{n}_i, \hat{v}_i, \hat{\delta}_i \hat{s}_i)$ and a generic ground truth annotation $(b_j, n_j, v_j, \delta_j)$, we define different variants of this Top-K evaluation measure considering different matching criteria to assess the ability of the model to predict next object interactions at different levels of granularity.

Long-term action anticipation: Methods are evaluated at the set of timestamps specified by the end of each annotated object interaction in a video V. Let $L_V^{(j)} = \{(n_z^{(j)}, v_z^{(j)})\}_{z=1}^{Z}$ be the ground truth annotation related to video V at time-stamp $t^{(j)}$ and let $\{\{(\hat{n}_{z,k}^{(j)}, \hat{v}_{z,k}^{(j)})\}_{z=1}^{Z}\}_{k=1}^{K}$ be the K predicted sequences of Z actions. The K predicted sequences will hence be evaluated using the edit distance metric. For a given k, this is obtained by evaluating the edit distance between a predicted sequence and the ground truth sequence of future actions. The edit distance

$$\Delta_E(\{(\hat{n}_{z,k}^{(j)}, \hat{v}_{z,k}^{(j)})\}_{z=1}^{Z}, \{(n_z^{(j)}, v_z^{(j)})\}_{z=1}^{Z})$$

is computed as the Damerau-Levenshtein distance over sequences of predictions of verbs, nouns and actions. The goal of this measure is to assess performance in a way which is robust to some error in the predicted order of future actions. A predicted verb/noun is considered "correct" if it matches the ground truth verb label at a specific time-step. The allowed operations to compute the edit distance are insertions, deletions, substitutions and transpositions of any two predicted actions. Following the "best of many" criterion, the K predictions are evaluated considering the smallest edit distance between the ground truth and any of the K predictions:

$$\Delta_E(\{\{(\hat{n}_{z,k}^{(j)}, \hat{v}_{z,k}^{(j)})\}_{z=1}^{Z}\}_{k=1}^{K}, \{(n_z^{(j)}, v_z^{(j)})\}_{z=1}^{Z}) =$$
$$\min_{k=1..K} \Delta_E(\{(\hat{n}_{z,k}^{(j)}, \hat{v}_{z,k}^{(j)})\}_{z=1}^{Z}, \{(n_z^{(j)}, v_z^{(j)})\}_{z=1}^{Z})$$

Note that we consider edit distance over simple accuracy based measures. Treating predictions for each future time-step independently and calculating accuracy does not account for the sequential nature of the prediction task where the order of predictions is important. We evaluate each metric independently for verbs, nouns and actions (verb and noun together). We report edit distance at $Z = 20$ (ED@20) and use $K = 5$ in our experiments. We select $Z = 20$ as baselines begin to predict actions at random for higher values of Z.

5) Baselines: Locomotion prediction: We make use of the method by Park et al. [111] for a baseline algorithm. The method models the trajectory prediction function of the following equation:

$$\mathcal{X} = \begin{bmatrix} \mathbf{x}_{t+1} & \cdots & \mathbf{x}_{t+F} \end{bmatrix}^{\mathsf{T}} = f(\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}; \mathcal{I}), \quad (8)$$

where $\mathcal{X}$ is the future trajectory, T and F are the past and future time horizons, respectively, $\mathbf{x}_t$ is the point on the trajectory at time t, and $\mathcal{I}$ is the egocentric image at time t. With an assumption that the person walks over a major plane (e.g., ground plane), we represent the trajectory in a 2D plane, i.e., $\mathbf{x}_t \in \mathbb{R}^2$.

We use KNN classification with CNN image encoding, i.e.,

$$\{\mathcal{X}\} = KNN\left(\{\phi(\mathcal{I}_i)\}, \phi(\mathcal{I})\right) \quad (9)$$

where $KNN(A, B)$ finds the K nearest neighbor of B given the set A, and $\phi(\mathcal{I}) \in \mathbb{R}^n$ is a function that extracts the image feature of $\mathcal{I}$. We use the AlexNet image feature extractor for ϕ.

Notably, the baseline algorithm leverages a polar coordinate system to represent the trajectory, i.e., $\mathbf{X}_j^{2D} = \begin{bmatrix} r_j & \theta_j \end{bmatrix}^{\mathsf{T}}$ is a 2D trajectory on the ground plane where r_i and θ_i are the polar coordinates of the trajectory represented in the egocentric coordinate system, i.e., distance (radial) and direction (angle) with respect to the person's feet location:

$$\mathbf{X}_j^{2D} = \texttt{cart2polar}(\mathbf{r}_1^{\mathsf{T}}\mathbf{X}_j, \mathbf{r}_2^{\mathsf{T}}\mathbf{X}_j) \quad (10)$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the two spanning vectors of the ground plane that are aligned with the rotation matrix $\mathbf{R}_t$. $\mathbf{r}_1$ is the facing direction and $\mathbf{r}_2$ is lateral direction. Both are perpendicular to the ground plane normal $\mathbf{n}$. `cart2polar` is a coordinate transform from cartesian to polar coordinates.

Hands movements prediction: The proposed future hand movement prediction task can be factorized as a regression problem. To address this task, we adopt a baseline that utilizes the I3D network as the backbone to extract the spatial-temporal video representations of the input video sequence, and then use a linear mapping function as the regressor to predict the future keyframe hand positions. We adopt the smoother l1 loss as the objective function:

$$L_h = \begin{cases} 0.5 * w * (h - \hat{h})^2/\beta, & \text{if } |h - \hat{h}| < \beta \\ w * (|h - \hat{h}| - 0.5 * \beta), & \text{otherwise} \end{cases} \quad (11)$$

where $h \in R^{20}$ is a vector that represents the x,y coordinates of both left and right hands in the aforementioned five future key frames. If the hand is not observed in the keyframe, we pad 0 into the $\hat{h}$, and adopt a binary mask w to prevent the gradients propagation of these unobserved instances.

Short-term object interaction anticipation: The baseline includes two main components. A Faster R-CNN object detector [122] is used to detect next active objects in the last frame of the input video clip processed at full resolution. A SlowFast 3D CNN [46] is hence used to predict a verb label and a time to action for each predicted object. This is done by obtained a fixed-length representation of each object through ROI pooling [122]. Two linear layers are hence used to predict a probability distribution over verbs and a positive quantity for time to contact prediction respectively. Verb probability distributions are obtained using a softmax layer, whereas a softplus activation is used for time to contact prediction to make sure that the prediction is a positive number. The final output of the model is obtained by attaching the predicted verb and time to contact to each detected next active object. The noun label and confidence scores are copied from the output of the Faster R-CNN component.

Long-term action anticipation: The goal of the baseline model is to take as input a trimmed video of arbitrary length, and predict N different plausible sequences of future actions. The baseline models thus consist of three components: (1) the encoder backbone for obtaining clip level features, (2) the aggregation module for combining the obtained features from different clips, and (3) the decoder network for decoding the plausible sequences of future actions. For encoder backbones, we consider state of the art video recognition networks from both convolutional model, namely, SlowFast [46] and the newly proposed video transformer models, namely, MViT [41]. For aggregation module, we experiment with simple concatenation operators that concatenates the obtained clip features from multiple input clips as well as transformer based self-attention modules. For the decoder networks we consider the following options:

- No Change: A simple recognition baseline that assumes no future change in the current action and simply predicts the *currently* observed action as a duplicated static future sequence for Z steps.
- MultiHead: This model trains Z independent heads in parallel, one for each future time step. The final sequence is simply the conjoined predicted actions of each head.

Finally, to generate N plausible future sequences for constructing multimodal baselines, we simply sample the predicted future action distribution N times.

6) Results: Locomotion prediction: We evaluate the KNN based baseline algorithm by measuring mean trajectory error (K-MTE) and probability of correct trajectory (PCT) given an error tolerance. The trajectory length ranges from 7 to 15 seconds (70-150 points in a trajectory given 10 FPS). Our baseline results are reported in Table VIIIa).

Hands movements prediction: Table VIIIb) reports the results of our baseline as well as the results obtained by the winning method of the ECCV 2022 challenge competition [19]. Regarding the results obtained by our baseline, it is worth noting that predicting hand positions on contact frame is more challenging than on other key frames. This is because, by the definition of contact frame and pre-condition frame, the anticipation temporal footprint of contact frame is larger than other key frames. Qualitative results of our baseline method are reported in Fig. 28. Notably, the model can make reasonable predictions on future

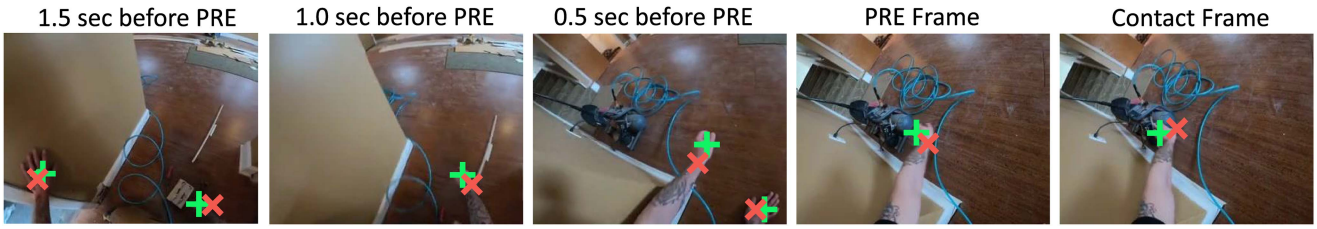


Fig. 28. Qualitative examples of future hands movements prediction using the proposed baseline. The ground truth hands positions are plotted as green crosses, while the predicted hands positions are plotted as red crosses.

Fig. 29. Qualitative examples of short-term object interaction anticipation using the proposed baseline. The numbers in brackets represent the confidence scores associated to the predictions. The ground truth next-active object is highlighted using a dashed red line, whereas model predictions are reported in blue solid lines.

TABLE VIII
FORECASTING RESULTS

	1-MTE	3-MTE	$\text{PCT}_{\epsilon=1m}$	$\text{PCT}_{\epsilon=3m}$
Ego4D Baseline	7.66m	5.54m	0.16	0.40

a) Locomotion prediction results.

	Left Hand		Right Hand	
	M.Disp.↓	C.Disp.↓	M.Disp.↓	C.Disp.↓
Ego4D Baseline	52.98	56.37	53.68	56.17
ECCV22 [19]	43.85	53.33	46.25	53.37

b) Hands movements prediction results.

	Noun	Noun+Verb	Noun+TTC	Overall
Ego4D Baseline	20.45	6.78	6.17	2.45
ECCV22 [19]	24.60	9.19	7.64	3.40

c) Short-term object interaction anticipation results.

	ED@(Z=20)		
	Verb	Noun	Action
Ego4D Baseline	0.739	0.780	0.943
ECCV22 [93]	0.741	0.739	0.930

d) Long-term action anticipation.

hand positions. However, the model is more likely to fail when there is drastic embodied motions.

Short-term object interaction anticipation: The baseline results are reported in Table VIIIc). The table also includes the results obtained by the winning method of the ECCV 2022 challenge competition for this task [19]. Fig. 29 reports some qualitative examples of the baseline. The model is sometimes able to detect the next active objects and predict suitable verbs and TTCs, but performance tends to be limited especially in complex scenarios.

Long-term action anticipation: Table VIIId) reports the baseline results together with the ones obtained by the winning method of the ECCV 2022 challenge competition [93]. Fig. 30 shows some qualitative results of our baseline. In each row, the ground truth future actions are shown along with the predictions from our model (for 5 time-steps). Correct predictions are highlighted in green, while valid actions that are incorrectly ordered (or paritally correct) are highlighted in blue. Note that though not perfectly aligned, incorrectly ordered sequences are given partial credit via the edit-distance metric.

7) Discussion: Locomotion prediction: The baseline quantitative results on the locomotion prediction task imply that the visual cues, e.g., side walk, obstacles, and road, in egocentric images are highly indicative of future movement. However, the baseline method that encodes the visual semantics of an image with a global feature is not detailed enough to model complex walking movement, e.g., avoiding pedestrians. This opens an opportunity for challenge participants to incorporate a fine-grained visual representation.

Hands movements prediction: Our baseline model for future hands movements prediction suffers from the drastic head movements in egocentric video and the stochastic nature of future forecasting. We speculate that explicitly modeling the head movements and next-active objects may complement the video representations for predicting future hands movements.

Short-term object interaction anticipation: The key challenges are likely due to the uncertain nature of future predictions as well as to the inability of the object detector to correctly detect next active objects and ignore the others. Nevertheless, the proposed baseline, even if simple, greatly improves over a

Fig. 30. Long term action anticipation - qualitative results. Actions in green represent correct predictions (correct action, at the correct position). Actions in blue represent incorrect ordering of valid actions. Our edit-distance metric accounts for both cases.

combination of an object detector and a random prediction of verbs and time to contact quantities. This suggests that methods can learn to analyze the input video in order to make reasonable predictions about the future.

Long-term action anticipation: We compared the performance of our baseline models when pretrained *only* on Kinetics-400 action recognition (as opposed to further fine-tuning on Ego4D action recognition). All models benefit greatly from training on Ego4D data because there is a large domain gap between Kinetics and Ego4D both in terms of visuals (third-person vs. egocentric viewpoint) and the diversity of activities they contain. Our transformer aggregation modules aggregate information across a larger temporal history controlled by the number of input clips to the model. Performance increases as more context information is provided to the model, however this increase comes at the cost of memory consumption. As expected, it is far easier to anticipate actions that occur immediately next, which gets more difficult as Z increases, and steadily plateaus. We evaluate the best of $K = 5$ predictions to arrive at our final results. To generate the K predictions, we sample each classifier head independently, however there are several methods to improve this including heuristic search algorithms (like beam search). Ideally, the multi-modal nature of future prediction should be accounted for in the model design itself. Moreover, decoder models that take into account the sequential nature during inference should be considered. These include transformer based decoders that are popular in recent language models (e.g., BERT, GPT). This is an important future direction of research.

VII. Annotation

Next we summarize the scope and process of the crowdsourced annotations, and explain how we derived taxonomies for objects, actions, and moments using the narrations.

A. Crowdsourced Annotations

To enable training and evaluation for this wide set of benchmarks, rich annotations accompany the Ego4D data. Beyond the 3.85M narration sentences discussed in Section IV, each Ego4D benchmark is also powered with a range of annotations. As a recap, for Episodic Memory, we label curated portions of the dataset with free-form natural language queries and temporal response windows (NLQ); temporal localizations of high-level events identified from a provided taxonomy (MQ); and visual crops with bounding box annotations of object occurrences as a response to sets of object queries (VQ). The Forecasting and Hands+Objects benchmarks benefit from bounding box annotations and labels for hands and objects captured at object transition points. Ego4D is also labeled for our Audio-Visual Diaraization and Social Interaction benchmarks, with temporal segments of voice activations, speech transcriptions (AVD), and attention cues individualized to a range of research subjects across diverse social settings (Social). See Table IX for the volume of these annotations across the existing Ego4D dataset. Please note that these annotations will be increased in a forthcoming update to the Ego4D dataset and these numbers are listed as well.

TABLE IX
Amount of Annotated Data for Each Benchmark

Benchmark	V1 Num hours	V2 Num Hours
EM VQ-2D	432.9	432.9
EM VQ-3D	13	13
EM Moments	328.7	328.7
EM NLQ	227.1	357.7
Hands+Obj.	196.2	196.2
Forecasting	110.5	110.5
AVD	47.7	47.7
Social	47.7	47.7

EM refers to Episodic Memory and its related sub-tasks. AVS refers to Audio-Visual Diarization. V1 refers to the current Ego4D dataset and basis of all baseline results reported in this paper. V2 refers to the additional annotations released in 2022. All 3,670 hours of video have narrations and features.

Annotations of this scale on video data are the product of an enormous team effort, involving the development of annotation guidelines, annotation pilots, tooling adjustments, trainings, quality control, selecting and curating clips to be labeled, and documenting annotations to support public use. Each Ego4D annotation was done by hand by over 300 professional annotators based in Lagos and Nairobi. Over a period of nearly one year,

these skilled annotators worked on this project full-time with a combined effort of more than 250,000 hours of human labor. We thank these hard working individuals for their contributions and partnership.

B. Taxonomies

We perform text mining over Ego4D narrations (see Section IV) to create various data-driven taxonomies for benchmark tasks.

Verb and Noun Taxonomy: In total the raw narrations describe the Ego4D video using 1,772 unique verbs and 4,336 unique nouns. Following ideas from [28], we leverage the narrations data to construct a taxonomy over the actions and objects that appear in the video, as follows. We use a part-of-speech (POS) tagger and dependency parser to identify verbs and nouns from each narrated action. We use an ensemble of parser models from the Spacy [59] toolkit to do this. Given a natural language narration, we first identify verbs using their POS tag. Then using the dependency tree, we identify all direct objects of the verb. To ensure verbs and nouns are accurately parsed, we adopt several heuristics: Parsed verbs are split into multiple senses (e.g., "turn" is split into "turn-on", "turn-off" and "turn-over"); compound nouns are decomposed into a root noun coupled with a modifier to ensure the noun taxonomy is unambiguous (e.g., modifier "egg" and root noun "shell" in "egg shell"); collective nouns are mapped to their main entity (e.g,. "piece of cheese" $\rightarrow$ "cheese"). Finally, we manually cluster the verbs and nouns to avoid redundancy in the taxonomy (e.g., "cut", "chop", "slice" are all mapped to the verb cluster "cut").

The resulting taxonomy consists of a set of 115 verbs ($\mathcal{V}$) and a set of 478 nouns ($\mathcal{N}$). Fig. 18 shows the distribution of verbs and nouns in a set of video data annotated with the taxonomy. This taxonomy is used for the forecasting tasks in Section VI-E.

Moments Activity Taxonomy: We next devise a taxonomy for the moments. Recall that moments are meant to be higher-level events, compared to the atomic actions used in forecasting. To that end, we use the *summary* narrations collected for five-minute clip segments, as they capture higher-level events and activities that are suitable for the moments retrieval task. This is in contrast to the verb-noun taxonomy that is sourced from individual narrations for each atomic action. The taxonomy was created as follows. First, each summary narration was encoded into a feature vector using a pre-trained BERT [33] language model, and then concatenated with the word embeddings for the main verb and noun extracted from the summary. These summaries were then clustered into groups, and then labels were manually assigned to groups based on the coherent activities they described.

Note that this process was done independently for a set of scenarios that we selected based on how frequently they occur in the dataset, the diversity of activities they represent, and how likely they contain high-level, event-like activities. For example videos that primarily involve a single activity like "driving" are not interesting categories in this context, whereas "household cleaning" contains several different activities that are shared across other indoor tasks, making it an appropriate scenario. In total, we select videos from five scenarios to create our moments taxonomy: Cooking, Cleaning, Shopping, Handyman, Farmer/Gardener. Fig. 16 shows the moments taxonomy.

VIII. Societal Impact

Our contribution can positively impact video understanding. It offers the research community a large-scale resource captured with rigorous privacy and ethics standards together with a diversity of subjects, and the benchmarks will promote reproducible technical advances. More broadly, egocentric perception has the potential to positively impact society in many application domains, including assistive technology, education, fitness, entertainment and gaming, eldercare, robotics, and augmented reality.

Nonetheless, future research in this area must guard against the potential negative societal impact if technology for egocentric vision were misused.

First, there are risks surrounding privacy. As we begin to see a proliferation of wearable cameras in public spaces, producers of these wearable devices will need to develop and implement protocols for notice and consent regarding the collection of data in public spaces, as well as user controls for how such data may be used, stored, and shared with any third parties. Similarly, models that may be used to transcribe speech or perform other tasks related to footage should include robust user controls such as the ability to remove or obscure personal data or sensitive content. Note that for all our audio-visual and social benchmarking work, the data used has full consent from the participants in the video, i.e., to use their unblurred faces and audio of their conversation. To date, the research community has lacked any large-scale data resource with which to study these kinds of problems; Ego4D will help the community to consider new solutions while leveraging real-world, diverse data that respects the privacy protocols of different countries. Furthermore, the Ego4D data is available only for users who sign a license that enumerates the allowable uses of the data, which is intended to hinder potential negative applications.

Second, there is a risk that our large-scale collection could inspire future collection efforts without the same level of care or attention to the privacy and ethical concerns as were taken in Ego4D. To mitigate this risk, we have aimed to be comprehensive in our descriptions of all parts of our procedures, and we include our best practices recommendations while publicly disseminating the results of the project.

Finally, despite our best efforts as discussed in this paper, there are still some imbalances in the dataset. For example, the data from Rwanda is relatively small, and though 74 cities represents a leap in coverage, they do not capture all possible demographics. We acknowledge that no matter how far one goes, full global coverage of daily life activity is elusive. Still, we can mitigate this risk by continuing to grow global collaborations with researchers and participants in underrepresented areas.

IX. Discussion and Future Directions

Like other dataset papers before it in computer vision, this paper on Ego4D has three main elements (1) the dataset and annotations (2) benchmark tasks and evaluation metrics (3) baseline implementations of models for solving the various

tasks. History tells us that the importance of these three elements follows the same order: the dataset is most important, the benchmarks less so, and least of all the model implementations. For example, for ImageNet [32], the most impactful computer vision dataset in the last two decades, what mattered most was the data, because when it was used for training AlexNet [71], it resulted in performance significantly higher than that of more traditional computer vision approaches based on hand-designed features like HOG [25], launching the deep learning revolution.

Who now remembers the baseline implementations in the papers that introduced once famous datasets like Caltech 101 [45], PASCAL [39], Kinetics [16] and so on? In an era of rapid progress, models become obsolete quickly. Ego4D has been around for a very short time. It was published at CVPR 2022 in June and already in the two benchmark competitions held at CVPR 2022 and ECCV 2022 we are seeing that models proposed in the original Ego4D paper are being superseded. This is how it should be.

New benchmark tasks will be added, and old ones will be fine-tuned. Recall that benchmarks in the area of object detection and segmentation evolved significantly from PASCAL [39] in 2005 to COCO in 2014 [85], by creating better metrics for evaluation. We expect and welcome further fine-tuning by the computer vision community of the Ego4D benchmarks and metrics for evaluations, as well as the addition of totally new ones, leveraging the size and scope of Ego4D.

The data is the most valuable element, because in addition to the original benchmark tasks, the research community will find novel ways to exploit it that we, the designers of the dataset, have not yet conceived.

Already one such application has emerged that is worth highlighting: training robotics models. Robotics is about connecting perception to action. Ego4D is an ideal dataset, because the egocentric view captures the image perceived by the agent, and then as the agent acts in the world, locomoting and navigating while avoiding collisions, or manipulating the state of some object with his/her hands, the consequences of the action are also revealed in the video. Thus both the input and output of the agent can be identified in the egocentric video stream. As the robotics community embraces machine learning more extensively, imitation learning has come to the fore. The Ego4D dataset supports multiple ways of making it easier for a robot to do a task given examples of a human performing it: (1) pre-training visual representations [103], [116], [143]. Egocentric data is more representative than, say, ImageNet, as the "eye of the beholder" is also the "eye of the actor" (2) learning affordances and rewards [87], [89]. In an egocentric video stream, there are many examples of how to open cabinet drawers and refrigerator doors, how to pick up coffee mugs, or how to hold and cut with a knife; (3) raw trajectories for training a robotic skill, e.g. how to walk through a room avoiding obstacles, or sweeping the floor, in addition to the object manipulation examples above.

Another major application area is at the interface of language and video. There are multiple datasets which have images with associated text, and these have been used to train models such as CLIP [115]. However, when we consider for example YouTube videos, typically the annotations are at the level of meta-data and more abstract e.g. "soccer match between England and France". In contrast, the narrations of the Ego4D video stream are much more fine-grained and at the level of specific actions performed on objects. These may prove to be very valuable for training multimodal language-video models [82].

What next? There are incremental updates that have been already been made to Ego4D v1 by collecting more data, annotating more of it, and linking to auxiliary signals such as IMUs, gaze, 3D scans, and the like; the expanded Ego4D v2 (released in early 2023) is the result and it will help train better models.

We have also initiated a significant novel data collection effort, "Ego-Exo", where we collect one or more additional "exocentric" video streams simultaneously and synchronized with the egocentric video. These exocentric views correspond to a third-person view of the activity of the ego-camera wearer. For example, for a camera wearer breaking an egg to make an omelet, the egocentric video will show a high resolution view of the hands, the egg, and the vessel to collect the egg yolk/whites while the exocentric view will capture the full body as well as the scene context of the kitchen. Our conjecture is that these two streams carry correlated and complementary information for analyzing the activity being performed by the human agent. Past experience on many computer vision problems suggests that there is a value in both "coarse" and "fine" views—scene context as well as specific objects. Inferring body pose is easiest from the exocentric view, while the interaction between the hands and the object being manipulated is better captured in the egocentric view. In the spirit of Ego4D, we will be collecting data in a wide variety of geographical locations with the goal of capturing the diversity in how the same activity is performed in different settings.

X. 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.

Acknowledgments

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, managers, or working group leads in the project. Appendices of [57] detail the contributions of individual authors for the various benchmarks. 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 deidentification 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 - CUPE64118002540007. Bristol is supported by UKRIEngineering and Physical Sciences Research Council (EPSRC) Doctoral Training Program (DTP), EPSRC Fellowship UMPIRE (EP/T004991/1). KAUST issupported by the KAUST Office of Sponsored Research through the Visual Computing Center (VCC) funding. National University of Singapore is supported by Mike Shous Start-Up Grant. Georgia Tech issupported in part by NSF 2033413 and NIH R01MH114999

Authors' Affiliations

Kristen Grauman, Tushar Nagarajan, and Santhosh Kumar Ramakrishnan are with FAIR, Menlo Park, CA 94025 USA, and also with the University of Texas at Austin, Austin, TX 78712 USA (e-mail: grauman@cs.utexas.edu).

Andrew Westbury, Rohit Girdhar, Jackson Hamburger, Devansh Kukreja, Miguel Martin, Dhruv Batra, Akshay Erapalli, Christoph Feichtenhofer, Satwik Kottur, Yanghao Li, and Tullie Murrell are with FAIR, Menlo Park, CA 94025 USA.

Eugene Byrne is with FAIR, Menlo Park, CA 94025 USA, and also with Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Vincent Cartillier is with FAIR, Menlo Park, CA 94025 USA, and also with Georgia Tech, Atlanta, GA 30332 USA.

Zachary Chavis, Jayant Sharma, Tien Do, and Hyun Soo Park are with the University of Minnesota, Minneapolis, MN 55455 USA.

Antonino Furnari and Giovanni Maria Farinella are with the University of Catania, 95124 Catania, Italy.

Hao Jiang, Chao Li, and Minh Vo are with the Meta Reality Labs, Menlo Park, CA 94025 USA.

Miao Liu, Morrie Doulaty, James Hillis, Jáchym Kolář, Anurag Kumar, Leda Sari, Kiran Somasundaram, Christian Fuegen, Vamsi Krishna Ithapu, Richard Newcombe, and Mingfei Yan are with the Meta, NW1 3FG London, U.K.

Xingyu Liu, Sean Crane, Qichen Fu, Xindi Wu, and Kris Kitani are with the Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Ilija Radosavovic and Karttikeya Mangalam are with the UC Berkeley, Berkeley, CA 94720 USA.

Fiona Ryan, Wenqi Jia, and Audrey Southerland are with Georgia Tech, Atlanta, GA 30332 USA.

Michael Wray, Siddhant Bansal, Adriano Fragomeni, Jonathan Munro, Will Price, and Dima Damen are with the University of Bristol, BS8 1QU Bristol, U.K.

Mengmeng Xu, Chen Zhao, Merey Ramazanova, and Bernard Ghanem are with the King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia.

Eric Zhongcong Xu, Ruijie Tao, Yunyi Zhu, Haizhou Li, and Mike Zheng Shou are with the National University of Singapore, Singapore 119077.

Abrham Gebreselasie is with the Carnegie Mellon University Africa, Kigali BP 6150, Rwanda.

Cristina González, Paola Ruiz Puentes, and Pablo Arbeláez are with the Universidad de los Andes, Santiago 12455, Chile.

Xuhua Huang is with the Carnegie Mellon University, Pittsburgh, PA 15213 USA, and also with Meta, NW1 3FG London, U.K.

Yifei Huang, Zhenqiang Li, Takumi Nishiyasu, Yusuke Sugano, Takuma Yagi, and Yoichi Sato are with the University of Tokyo, Tokyo 113-8654, Japan.

Weslie Khoo, Yuchen Wang, Ziwei Zhao, and David Crandall are with the Indiana University, Bloomington, IN 47405 USA.

Federico Landini is with FAIR, Menlo Park, CA 94025 USA, and also with the Brno University of Technology, Brno 601 90, Czechia.

Raghava Modhugu and C. V. Jawahar are with the International Institute of Information Technology, Hyderabad, Hyderabad, Telangana 500032 USA.

Hanbyul Joo is with FAIR, Menlo Park, CA 94025 USA, and also with Seoul National University, Seoul 08826, South Korea.

Aude Oliva and Antonio Torralba are with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

James M. Rehg is with Georgia Tech, Atlanta, GA 30332 USA, and also with the University of Illinois Urbana-Champaign, Champaign, IL 61820 USA.

Jianbo Shi is with the University of Pennsylvania, Philadelphia, PA 19104 USA.

Lorenzo Torresani is with FAIR, Menlo Park, CA 94025 USA, and also with Dartmouth College, Hanover, NH 03755 USA.

Jitendra Malik is with FAIR, Menlo Park, CA 94025 USA, and also with UC Berkeley, Berkeley, CA 94720 USA.

REFERENCES

[1] 2nd International Ego4D Workshop, @ ECCV2022. [Online]. Available: https://ego4d-data.org/workshops/eccv22/

[2] Github repository of the ESPNet model Zoo. [Online]. Available: https://github.com/espnet/espnet_model_zoo

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.

[4] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018.

[5] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.

[6] X. A. Miró, "Robust speaker diarization for meetings," Ph.D. thesis, Universitat Politécnica de Catalunya, 2006.

[7] R. Arandjelović and A. Zisserman, "Objects that sound," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 451–466.

[8] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1949–1957.

[9] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Proc. IEEE 6th Int. Workshop Vis. Surveill. Conjunction ECCV*, Citeseer, 2006.

[10] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

[11] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "First-person action-object detection with EgoNet," in *Proc. Robot. Sci. Syst.*, 2017.

[12] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 441–456, Feb. 2018.

[13] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D object pose estimation using 3D object coordinates," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 536–551.

[14] H. Bredin et al., "Pyannote. audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 7124–7128.

[15] M. Cai, K. M. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Proc. Robot. Sci. Syst.*, 2016.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[17] S. Chaudhuri et al., "Ava-speech: A densely labeled dataset of speech activity in movies," 2018, *arXiv:1808.00606*.

[18] C. Chen et al., "Audio-visual embodied navigation," *Environment*, vol. 97, 2019, Art. no. 103.

[19] G. Chen et al., "InternVideo-Ego4D: A pack of champion solutions to Ego4D challenges," 2022, *arXiv:2211.09529*.

[20] E. Chong et al., "Detection of eye contact with deep neural networks is as accurate as human experts," *Nature Commun.*, vol. 11, no. 1, Dec. 2020, Art. no. 6386.

[21] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg, "Detecting attended visual targets in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, 2020, pp. 5395–5405.

[22] J. S. Chung et al., "In defence of metric learning for speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020.

[23] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: Speaker diarisation in the wild," 2020, *arXiv: 2007.01216*.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[26] D. Damen et al., "The EPIC-KITCHENS dataset: Collection, challenges and baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4125–4141, Nov. 2021.

[27] D. Damen et al., "Rescaling egocentric vision," *Int. J. Comput. Vis.*, vol. 130, pp. 33–55, 2022.

[28] D. Damen et al., "Scaling egocentric vision: The EPIC-KITCHENS dataset," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 753–771.

[29] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas, "You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[30] D. Damen, T. Leelasawassuk, and W. Mayol-Cuevas, "You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance," *Comput. Vis. Image Understanding*, vol. 149, pp. 98–112, 2016.

[31] A. G. Del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 65–76, Feb. 2017.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv: 1810.04805*.

[34] J. Donley et al., "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments," 2021, *arXiv:2107.04174*.

[35] B. Doosti, C. Chen, R. Vemulapalli, X. Jia, Y. Zhu, and B. Green, "Boosting image-based mutual gaze detection using pseudo 3D gaze," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 1273–1281.

[36] H. Doughty, I. Laptev, W. Mayol-Cuevas, and D. Damen, "Action modifiers: Learning from adverbs in instructional videos," 2019, *arXiv: 1912.06617*.

[37] M. Dunnhofer, A. Furnari, G. M. Farinella, and C. Micheloni, "Visual object tracking in first person vision," *Int. J. Comput. Vis.*, vol. 131, pp. 259–283, 2023.

[38] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *Proc. ACM SIGGRAPH Conf. Exhib. Comput. Graph. Interactive Techn.*, 2018.

[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[40] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.

[41] H. Fan et al., "Multiscale vision transformers," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 6824–6835.

[42] Y. Fang et al., "Dual attention guided gaze target detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11385–11394.

[43] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1226–1233.

[44] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2579–2586.

[45] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, Art. no. 178.

[46] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6201–6210.

[47] A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella, "Next-active-object prediction from egocentric videos," *J. Vis. Commun. Image Representation*, vol. 49, pp. 401–411, 2017.

[48] A. Furnari and G. Farinella, "Rolling-unrolling LSTMs for action anticipation from first-person video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4021–4036, Nov. 2021.

[49] V. Gabeur, P. H. Seo, A. Nagrani, C. Sun, K. Alahari, and C. Schmid, "AVATAR: Unconstrained audiovisual speech recognition," 2022, *arXiv:2206.07684*.

[50] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 36–54.

[51] R. Gao and K. Grauman, "2.5D visual sound," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 324–333.

[52] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3878–3887.

[53] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15490–15500.

[54] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, May 2018.

[55] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 13485–13495.

[56] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16081–16091.

[57] K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2022, pp. 18973–18990.

[58] C. Gu et al., "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6047–6056.

[59] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," 2020, doi: 10.5281/zenodo.1212303.

[60] G. Irie et al., "Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 3961–3964.

[61] P. Isola, J. J. Lim, and E. H. Adelson, "Discovering states and transformations in image collections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1383–1391.

[62] K. Iwano, T. Yoshinaga, S. Tamura, and S. Furui, "Audio-visual speech recognition using lip information extracted from side-face images," *EURASIP J. Audio Speech Music Process.*, vol. 2007, pp. 1–9, 2007.

[63] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-C. Zhu, "A multi-view dataset for learning multi-agent multi-task activities," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 767–786.

[64] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3D body pose from egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3501–3509.

[65] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv: 1705.06950*.

[66] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "EPIC-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5492–5501.

[67] P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6911–6920.

[68] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Commun.*, vol. 38, no. 1/2, pp. 19–28, 2002.

[69] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal Communication in Human Interaction*, 8th ed.Wadsworth, OH, USA: Cengage Learning, 2014.

[70] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 706–715.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[72] F. D. la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey, "Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-22, 2009.

[73] L. Lacheze, Y. Guo, R. Benosman, B. Gas, and C. Couverture, "Audio/video fusion for objects recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2009, pp. 652–657.

[74] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1346–1353.

[75] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *Int. J. Comput. Vis.*, vol. 114, pp. 38–55, 2015.

[76] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion–A systematic study," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 443–455, Fourth Quarter 2012.

[77] C. Li and K. Kitani, "Model recommendation with virtual probes for ego-centric hand detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2013, pp. 2624–2631.

[78] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3216–3223.

[79] Y. Li, M. Liu, and J. Rehg, "In the eye of the beholder: Gaze and actions in first person video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6731–6747, Jun. 2023.

[80] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 619–635.

[81] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman, "Ego-Exo: Transferring visual representations from third-person to first-person videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6939–6949.

[82] K. Q. Lin et al., "Egocentric video-language pretraining," 2022, *arXiv:2206.01670*.

[83] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3889–3898.

[84] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[85] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[86] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–A new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6536–6545.

[87] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: Towards universal visual reward and representation via value-implicit pre-training," 2022, *arXiv:2210.00030*.

[88] J. Mai, C. Zhao, A. Hamdi, S. Giancola, and B. Ghanem, "Estimating more camera poses for ego-centric videos is essential for VQ3D," 2022, *arXiv:2211.10284*.

[89] P. Mandikal and K. Grauman, "DexVIP: Learning dexterous grasping with human hand pose priors from video," in *Proc. Conf. Robot Learn.*, 2022, pp. 651–661.

[90] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "Laeo-net: Revisiting people looking at each other in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3477–3485.

[91] M. J. Marín-Jiménez, A. Zisserman, M. Eichner, and V. Ferrari, "Detecting people looking at each other in videos," *Int. J. Comput. Vis.*, vol. 106, no. 3, pp. 282–296, 2014.

[92] M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari, "Here's looking at you, kid. Detecting people looking at each other in videos," in *Proc. Brit. Mach. Vis. Conf.*, 2011.

[93] E. V. Mascaro, H. Ahn, and D. Lee, Intention-conditioned long-term human egocentric action forecasting, Ego4D challenge 2022, 2022*arXiv:2207.12080*.

[94] I. McCowan et al., "The AMI meeting corpus," in *Proc. 5th Int. Conf. Methods Techn. Behav. Res.*, 2005, pp. 137–140.

[95] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2630–2640.

[96] K. Min, "Intel labs at Ego4D challenge 2022: A better baseline for audio-visual diarization," 2022, *arXiv:2210.07764*.

[97] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1160–1169.

[98] Y. Mitsuzumi, A. Nakazawa, and T. Nishida, "Deep eye contact detector: Robust eye contact bid detection using convolutional neural network," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

[99] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360° video," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 360–370.

[100] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8687–8696.

[101] T. Nagarajan and K. Grauman, "Attributes as operators: Factorizing unseen attribute-object compositions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 169–185.

[102] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017.

[103] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," 2022, *arXiv:2203.12601*.

[104] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6817–6826.

[105] L. Neumann, A. Zisserman, and A. Vedaldi, "Future event prediction: If and when," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2935–2943.

[106] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2Me: Inferring body pose in egocentric video via first and second person interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9887–9897.

[107] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[108] C. Northcutt, S. Zha, S. Lovegrove, and R. Newcombe, "EgoCom: A multi-person multi-modal egocentric communications dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6783–6793, Jun. 2023.

[109] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 639–658.

[110] C. Palmero et al., "Automatic mutual gaze detection in face-to-face dyadic interaction videos," *Measuring Behav.*, vol. 1, 2018.

[111] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4697–4705.

[112] H. S. Park, E. Jain, and Y. Sheikh, "3D social saliency from head-mounted cameras," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 422–430.

[113] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2847–2854.

[114] S. Purushwalkam, M. Nickel, A. Gupta, and M. Ranzato, "Task-driven modular networks for zero-shot compositional learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3593–3602.

[115] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[116] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," 2022, *arXiv:2210.03109*.

[117] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, and G. M. Farinella, "EGO-CH: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision," *Pattern Recognit. Lett.*, vol. 131, pp. 150–157, 2020.

[118] F. Ragusa, A. Furnari, S. Livatino, and G. M. Farinella, "The Meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1569–1578.

[119] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 199–207.

[120] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral imaging and autism," *IEEE Pervasive Comput.*, vol. 13, no. 2, pp. 84–87, Second Quarter 2014.

[121] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[122] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[123] I. Rodin, A. Furnari, D. Mavroedis, and G. M. Farinella, "Predicting the future from first person (egocentric) vision: A survey," *Comput. Vis. Image Understanding*, vol. 211, 2021, Art. no. 103252.

[124] J. Roth et al., "AVA active speaker: An audio-visual dataset for active speaker detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 4492–4496.

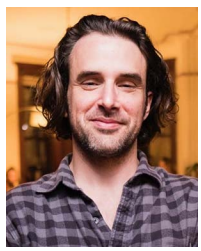
[125] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2730–2737.

[126] F. Sener et al., "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21064–21074.

[127] A. Senocak, T.-H. Oh, J. Kim, M. Yang, and I. S. Kweon, "Learning to localize sound sources in visual scenes: Analysis and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1605–1619, May 2021.
[128] D. Shan, J. Geng, M. Shu, and D. Fouhey, "Understanding human hands in contact at internet scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9866–9875.
[129] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," 2018, *arXiv: 1804.09626*.
[130] Silero Team, "Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," 2021. [Online]. Available: https://github.com/snakers4/silero-vad
[131] M. Silva, W. Ramos, J. Ferreira, F. Chamone, M. Campos, and E. R. Nascimento, "A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2383–2392.
[132] K. K. Singh, K. Fatahalian, and A. A. Efros, "KrishnaCam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.
[133] E. Spera, A. Furnari, S. Battiato, and G. M. Farinella, "Egocentric shopping cart localization," in *Proc. Int. Conf. Pattern Recognit.*, 2018, pp. 2277–2282.
[134] Y.-C. Su and K. Grauman, "Detecting engagement in egocentric video," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 454–471.
[135] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," 2021, *arXiv:2107.06592*.
[136] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 252–268.
[137] E. Tulving, "Episodic and semantic memory," in *Organization of Memory*, E. Tulving and W. Donaldson, Eds. Cambridge, MA, USA: Academic Press, 1972.
[138] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.
[139] X. Wang, A. Farhadi, and A. Gupta, "Actions $\sim$ transformations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2658–2667.
[140] X. Wang, L. Zhu, H. Wang, and Y. Yang, "Interactive prototype learning for egocentric action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8148–8157.
[141] X. Wang, L. Zhu, Y. Wu, and Y. Yang, "Symbiotic attention for egocentric action recognition with object-centric alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6605–6617, Jun. 2023.
[142] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," 2020, *arXiv: 2001.08740*.
[143] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," 2022, *arXiv:2203.06173*.
[144] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5288–5296.
[145] M. Xu, Y. Li, C.-Y. Fu, B. Ghanem, T. Xiang, and J.-M. Perez-Rua, "Where is my wallet? Modeling object proposal sets for egocentric visual query localization," 2022, *arXiv:2211.10528*.
[146] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2629–2638.
[147] H. Zhang, X. Cao, and R. Wang, "Audio visual attribute discovery for fine-grained object recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, Art. no. 924.
[148] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6543–6554.
[149] S. Zhang et al., "EgoBody: Human body shape and motion of interacting people from head-mounted devices," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 180–200.
[150] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D temporal adjacent networks for moment localization with natural language," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 12870–12877.
[151] C. Zhao, A. K. Thabet, and B. Ghanem, "Video self-stitching graph network for temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13658–13667.
[152] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 587–604.
[153] Y.-D. Zheng, G. Chen, J. Wang, T. Lu, and L. Wang, "Exploring state change capture of heterogeneous backbones @ Ego4D hands and objects challenge 2022," 2022, *arXiv:2211.08728*.
[154] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[155] Y. Zhou and T. L. Berg, "Temporal perception and prediction in egocentric video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 4498–4506.
[156] Y. Zhou and T. L. Berg, "Learning temporal transformations from time-lapse videos," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 262–277.
[157] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Automat. Comput.*, vol. 18, pp. 351–376, 2021.

Kristen Grauman (Fellow, IEEE) is a Professor in the Department of Computer Science at the University of Texas at Austin and a Research Director in FAIR with Meta. Before joining UT Austin in 2007, she received her PhD at MIT. She is an AAAI fellow, Sloan fellow, AAAS fellow, and a recipient of the PAMI Young Researcher Award and Computers and Thought Award. She and her collaborators have been recognized with several Best Paper awards in computer vision, including a 2011 Marr Prize and a 2017 Helmholtz Prize. She served as a program chair of CVPR 2015, NeurIPS 2018, and ICCV 2023, and as an associate editor-in-chief for *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Andrew Westbury is a research program manager with FAIR. He supports implementation of fundamental AI research projects focused on First-Person Perception and Embodied AI.



Eugene Byrne had the honor to serve as technical lead with FAIR for Ego4D. His engineering career spans 20 years across multimodal understanding, quantitative finance, virtual assistants and robotics. He's currently the CTO of Whysaurus focused on collective argumentation and consulting with companies in the agents and edtech space.

Vincent Cartillier is a PostDoc researcher with GeorgiaTech. His expertise is on 3D computer vision.

Zachary Chavis is currently working toward the PhD degree with the University of Minnesota advised by Dr. Hyun Soo Park and Dr. Stephen J. Guy focusing on applications of human motion for robotics, computer vision, and graphics.

Antonino Furnari (Senior Member, IEEE) received the PhD degree in mathematics and computer science from the University of Catania, in 2017. He is a tenure-track assistant professor with the University of Catania. He spent time as a visiting researcher with the University of Texas at Austin and with the University of Bristol. He has been working on First Person (Egocentric) Computer Vision since 2014 and he is part of the EPIC-KITCHENS and EGO4D teams. His research focuses on understanding human activity and future intent from egocentric video.

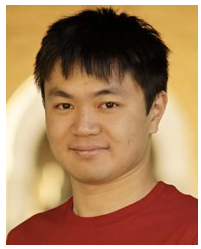
Rohit Girdhar received the MS and PhD degrees in robotics from Carnegie Mellon University, where he worked on learning from and understanding videos. He is a research scientist with the GenAI Research group, Meta. His current research focuses on understanding and generating multimodal data, using minimal human supervision. He was previously part of the Facebook AI Research (FAIR) group with Meta, and has spent time with DeepMind, Adobe and Facebook as an intern. His research has won multiple international challenges, and has been recognized through a Best Paper (Finalist) Award at CVPR'22, Best Paper Award at ICCV'19 HVU Workshop, Siebel Scholarship at CMU, and a Gold Medal and Research Award for undergraduate research at IIIT Hyderabad. He regularly serves on academic committees, such as on the Ego4D board, as an area chair for NeurIPS, CVPR, and ECCV, and has organized multiple workshops at premier computer vision conferences.

Jackson Hamburger is the CTO & co-founder with Sensorum, an AI for healthcare company. For the Ego4D initiative, he was the lead research engineer during the first phase. His interests span from fundamental research on novel AI methods to their applications in commercial and clinical contexts to enable never-before-possible technological products.

Hao Jiang received the PhD degree in computer science from Simon Fraser University in 2006. He is a research scientist with Facebook Reality Labs. From 2007 to 2017, he was an assistant professor and then a tenured associate professor with the Computer Science Department, Boston College. His research spans human pose, tracking, action understanding, 3D computer vision, egocentric vision and deep learning. From 2017 to 2020, he was a principal researcher with Microsoft Cloud and AI working on real-time 4D computer vision.

Devansh Kukreja received the undergraduate degree from Carnegie Mellon University. He is a SWE with FAIR. He moved from a fullstack background into AI Research. His research interests include perception (particularly ego-centric), visualization, and embodied AI.

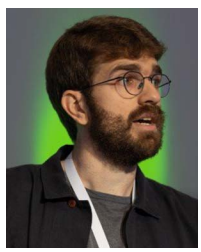
Miao Liu received the PhD degree in robotics from Georgia Tech in 2022 for his research on egocentric vision and visual attention. He is a reserach scientist with GenAI, Meta. He was a visiting researcher with ETH Zurich and Max Planck Institute and an intern with Facebook Reality Lab. He won the Best Student Paper Prize at BMVC 2022. His currect research focuses on multimodal large language model.



Xingyu Liu is a postdoc with Carnegie Mellon University.

Miguel Martin is a SWE with Meta FAIR.

Tushar Nagarajan received the PhD degree in computer science from UT Austin in 2022 for his work on egocentric video understanding and embodied AI. He is a research scientist with FAIR, Meta. His current research focuses on video and language understanding. During his PhD, he was a a visiting researcher with FAIR and has spent time with IBM Research as an intern. He is a running organizer of the EgoVis (Ego4D) workshop.



Ilija Radosavovic is currently working toward the PhD degree with UC Berkeley, advised by Jitendra Malik. His research focuses on robot learning from Internet videos. Prior to UC Berkeley, he was a research engineer with Facebook AI Research working with Piotr Dollar, Ross Girshick, and Kaiming He. Ilija is a recipient of the PAMI Mark Everingham Award (2021), and his work has been deployed across the industry and adopted by major corporations, including Facebook, Intel, and Tesla.

Santhosh Kumar Ramakrishnan completed his PhD thesis on predictive scene representations for embodied AI with UT Austin under the guidance of Dr. Kristen Grauman. His work focuses primarily on computer vision and representation learning for robotics and egocentric video understanding.

Fiona Ryan is currently working toward the PhD degree in computer science with the Georgia Institute of Technology. Her research focuses on computer vision systems for understanding human behavior, particularly in social contexts. She is a recipient of the NSF Graduate Research Fellowship.

Jayant Sharma leads the Perception team with Playtag-a startup founded to offer behavior analytics as a service, based on human tracking solutions in complex environments. As a graduate student, he headed University of Minnesota's efforts at data collection and processing for Ego4D.

Michael Wray is a lecturer/assistant professor of computer vision with the Department of Computer Science, the University of Bristol. His research interests are in multi-modal video understanding, particularly for egocentric videos–focusing on how both vision and language can be tied together towards tasks, such as cross-modal retrieval, grounding, and captioning.

Mengmeng Xu (Frost) is an AI research scientist with Meta. He is interested in video understanding and generation.

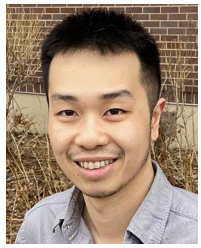
Eric Zhongcong Xu received the bachelor's degree from Tongji University in 2020. He is currently working toward the final-year PhD degree with the National University of Singapore, supervised by Asst. Prof. Mike Shou. His research interests include diffusion model, video generation, and 3D computer vision.

Chen Zhao received the PhD degree from Peking University (PKU) in 2016, and studied in University of Washington from 2012 to 2013. She is a research scientist with the King Abdullah University of Science and Technology (KAUST). Her research interests include computer vision, deep learning, image/video processing and compression, with a focus on video understanding. She received several awards such as the Best Paper Award in CVPR workshop 2023, and the Best Paper Award in NCMT 2015, and First Prize of the Qualcomm Innovation Fellowship Contest (QInF) 2012.

Siddhant Bansal received the master's degree from CVIT, IIIT Hyderabad, working with Prof. C.V. Jawahar and Prof. Chetan Arora. He is currently working toward the first-year PhD degree with the University of Bristol working with Prof. Dima Damen. His research interests are in devising learning-based methods for understanding and exploring various aspects of first-person (egocentric) vision.

Dhruv Batra is an associate professor with the School of Interactive Computing, Georgia Tech and a research director with the Fundamental AI Research (FAIR) team, Meta. His research interests lie at the intersection of machine learning, computer vision, natural language processing, and AI. He is a recipient of a number of awards, including PECASE, ONR YIP, NSF CAREER, and several best paper awards and nominations.

Sean Crane is a research associate with the Robotics Institute, Carnegie Mellon University.



Tien Do is currently working toward the PhD degree with the University of Minnesota.

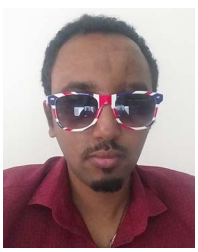
Morrie Doulaty received the PhD degree in computer science, speech technology from the University of Sheffield in 2016 working on domain adaptation of automatic speech recognition models. He is a software engineer with Meta since 2020. Prior to joining Meta he was a senior applied scientist with Microsoft from 2017 to 2020.

Akshay Erapalli is a Sr. technical program manager with Zoox working in the AI research space. While he was with Meta, he supported the implementation of fundamental AI research projects focused on First-Person Perception and Embodied AI.

Christoph Feichtenhofer received the BSc, MSc, and PhD degrees (all with distinction) in computer science from TU Graz in 2011, 2013 and 2017, respectively and spent time as a visiting researcher with York University, Toronto, as well as the University of Oxford. He is a research scientist manager with Meta AI (FAIR). He is a recipient of the PAMI Young Researcher Award, the DOC Fellowship of the Austrian Academy of Sciences, and the Award of Excellence for outstanding doctoral theses in Austria. His main areas of research include the development of effective representations for image and video understanding.

Adriano Fragomeni is currently working toward the fourth-year PhD degree in computer vision with the University of Bristol supervised by Professor Dima Damen and Dr. Michael Wray. His research focuses on multimodal video understanding, particularly in studying the interaction between vision and language in order to solve vision-language tasks (i.e., video retrieval).

Qichen Fu is a researcher in the master's of Robotics (MSR) with the Robotics Institute of Carnegie Mellon University, supervised by Professor Kris Kitani. His research interests include understanding human activity, reconstructing 3D objects/scenes, learning to interact with the world, and multimodal machine learning. In CMU, he focused on hand-object interaction.

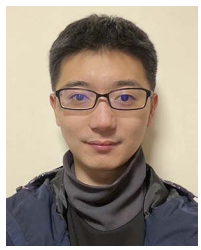


Abrham Gebreselasie is a research associate with the CMU-Africa campus working with Kris Kitani.

Cristina González received the graduation degree in biomedical engineering and computer science from Universidad de los Andes in 2021. She is currently working toward the PhD degree in engineering with the Center for Research and Formation in Artificial Intelligence (CINFONIA), supervised by Prof. Pablo Arbeláez with Universidad de Los Andes in Bogotá, Colombia, 2023, where she is a researcher. Her research interests include open visual recognition, natural language processing, biomedical image analysis, and egocentric vision.

James Hillis received the PhD degree in vision science from UC Berkeley on multi-sensory perception prior to a postdoctoral fellowship with the University of Pennsylvania on color perception, lectureship with the University of Glasgow where he focussed on social cognition. He is a research scientist with Meta Reality Labs where he focuses on development of sensor-to-display adaptive image-processing pipelines for AR.

Xuhua Huang is a researcher with the master's of science computer vision (MSCV) program with Carnegie Mellon University.



Yifei Huang received the PhD degree in information science and technology from the University of Tokyo in 2021. He is a special foreign researcher with the University of Tokyo. His current research focuses on video understanding and its applications in VR/AR.

Wenqi Jia is currently working toward the PhD degree in computer science with the Georgia Institute of Technology, supervised by James Rehg and co-advised by Danfei Xu. Her research focuses on human behavior understanding from egocentric vision and multi-modal learning.

Weslie Khoo received the BS degree in chemistry and biological chemistry from Nanyang Technological University and the PhD degree in food science from the Pennsylvania State University. He is a United States Department of Agriculture-NIFA postdoctoral fellow with Indiana University, advised by Prof. David Crandall. His research interests include using AI techniques for dietary assessments and robot-human interactions.

Jáchym Kolář received the PhD degree in artificial intelligence from the University of West Bohemia and the postdoc degree from LIMSI-CNRS in France as part of the Quaero project. He is a machine learning engineer with Meta, currently working with the AI Speech team in London. Prior to Meta, he was employed with Nuance Communications, where he worked on automatic medical speech transcription, and with Neuron Soundware, where his focus was on anomaly machine sound detection. His research interests include audio processing, spontaneous speech recognition, speech prosody, and speech translation.

Satwik Kottur received the doctorate and master's degrees from Carnegie Mellon University, with a focus on visual dialog agents. He is a research scientist working with FAIR, Meta. His research interests include egocentric multimodal reasoning and multimodal conversational agents. He has also spent time as a research intern with Google research and Snap research in the past.

Anurag Kumar received the undergraduate degree in electrical engineering from IIT Kanpur in 2013, and the PhD degree from Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA in 2018. He is currently a research lead and scientist with Reality Labs Research, Meta. His research interests include deep learning, audio and speech processing, multimodal learning, audio-visual scene understanding and generation and generative AI in multimodal domain. He currently serves on the *IEEE Audio and Acoustic Signal Processing* (AASP) Technical Committee.

Federico Landini received the Licentiate degree in computer science from the University of Buenos Aires, Argentina. He is currently working toward the PhD degree with the Brno University of Technology, Czech Republic. His research focuses on speaker diarization and machine learning applied to speech technologies.

Chao Li received the BE degree from Beijing Jiaotong University, the ME degree from Peking University, and the PhD degree in computer science from the University of Texas at Dallas. He is a research scientist with Meta Reality Labs Research. His current research focuses on computational imaging and egocentric vision.

Yanghao Li received the BS and MS degrees from Peking University. He is a research scientist with Apple. Before joining Apple, he was a research engineer with FAIR, Meta.

Zhenqiang Li received the PhD degree in information science and technology from the University of Tokyo in 2022. His research interests include human activity understanding and interpretable neural networks.

Karttikeya Mangalam received the PhD degree in long-term video and language understanding from the University of California, Berkeley advised by Prof. Jitendra Malik. He has since started a company focused on enhacing the phuman learning using the modern advances in multimodal LLMs.

Raghava Modhugu received the master's degree in computer science from IIIT Hyderabad with a specialisation in AI and computer vision under the guidance of Prof. C. V. Jawahar. His research interests are in the intersection of computer vision and robotics such as autonomous navigation, ego-centric vision with a focus on deep learning. He is currently working as senior lead engineer with comfort and driving assistance R&D team.

Jonathan Munro completed his PhD thesis on video domain adaptation from the University of Bristol in 2021. He was a member of the EPIC-Kitchens team, collecting the dataset and running the Domain Adaptation Challenge.

Tullie Murrell is the CEO and co-founder of Shaped, a company that's revolutionizing how businesses build behavioural driven applications with AI. Before that, he was an applied research scientist with Meta for several years focusing on video understanding and medical imaging research.

Takumi Nishiyasu received the MS degree in information science and technology from the University of Tokyo, in 2020. He is currently working toward the PhD degree in information science and technology with the Unversity of Tokyo supervised by Professor Yoichi Sato. His research interests include human activity understanding, gaze behavior analysis, and multimedia understanding.

Will Price received the PhD degree from the University of Bristol under the supervision of Prof. Dima Damen in 2021. His thesis focused on the role of time in video understanding. He participated in the collection of release of the EPIC-KITCHENS dataset.

Paola Ruiz Puentes is currently working toward the master's degree in Biomedical Engineering. She focuses on deep learning to address medical challenges. She has 4.5 years of experience in deep learning techniques for drug discovery, high-definition image analysis, and a holistic understanding of surgical scenes. Her primary research focuses on predicting the interactions of small molecules and disease-related proteins for drug repurposing and, additionally, on predicting the bioactivities of peptides to attack the increasing concern on antibiotic resistance.

Merey Ramazanova is currently working toward the PhD degree with the King Abdullah University of Science and Technology (KAUST), supervised by Professor Bernard Ghanem. Her current research focuses on multimodal video understanding.

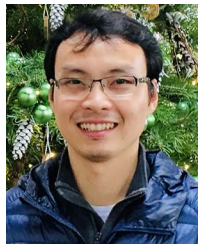
Leda Sari received the BSc and MSc degrees from Bogazici University, Türkiye and the PhD degree from the Department of Electrical and Computer Engineering, the University of Illinois Urbana-Champaign in 2021. She is a research scientist with Meta working on automatic speech recognition. Her research interests include speaker adaptation for automatic speech recognition, fairness for speech applications, and multi-modal automatic speech recognition.

Kiran Somasundaram received the doctorate and master's degrees from the University of Maryland, College Park. He is a lead software architect with Meta Reality Labs, Research. His research work focuses on building machine perception stacks for enabling all-day wearable AR/smart glasses.

Audrey Southerland received the bachelor's of science degree in psychology from Georgia Tech. She is a laboratory manager with the School of Interactive Computing at the Georgia Institute of Technology. She has overseen data collection, IRB protocol management, and data analysis for over a decade, working primarily with Dr. James Rehg and Dr. Agata Rozga. She has specialized in research with young toddlers, including those with developmental disabilities and autism.

Yusuke Sugano received the PhD degree in information science and technology from the University of Tokyo in 2010. He is an associate professor with the Institute of Industrial Science, The University of Tokyo. His research interests focus on computer vision and human-computer interaction. He was previously an associate professor with the Graduate School of Information Science and Technology, Osaka University, a postdoctoral researcher with Max Planck Institute for Informatics, and a project research associate at the Institute of Industrial Science, the University of Tokyo.

Ruijie Tao received the bachelor's degree from Soochow University in 2018 and the master's and PhD degrees from NUS in 2019 and 2023, respectively. He is a research fellow in National University of Singapore (NUS), supervised by Prof. Li Haizhou. His research interest audio-visual speech processing, includes speaker recognition, active speaker detection, self-supervised learning.



Minh Vo received the PhD degree from The Robotics Institute, Carnegie Mellon University. He is the head of Engineering with SpreeAI, a high-tech virtual try-on startup, where he oversees all products R&D. Before joining SpreeAI, he was a senior research scientist with Meta Reality Labs Research, where he worked on 3D perception and human sensing algorithms for Meta Aria glasses.

Yuchen Wang received the BE degree in software engineering from Tongji University in 2011, and the MS degree in computer science in 2017 from Indiana University, Bloomington, where he is currently working toward the PhD degree with the School of Informatics and Computing. His research interests include computer vision and robotics.

Xindi Wu is a researcher in the master's of science computer vision (MSCV) program with Carnegie Mellon University.

Takuma Yagi received the PhD degree in information science and technology from the University of Tokyo in 2022 and later worked as a project researcher with the Institute of Industrial Science, The University of Tokyo. He is a research scientist with the National Institute of Advanced Industrial Science and Technology (AIST) and a cooperative research fellow with the University of Tokyo. His research interests include human activity understanding, egocentric vision, and hand-object interaction understanding.

Ziwei Zhao received the BS degree in physics from the University of Science and Technology of China (USTC) and the MS degree in electrical engineering from Washington University in St. Louis (WUSTL). He is currently working toward the PhD degree in computer science with Indiana University, supervised by Prof. David Crandall. His research focuses on Egocentric Vision.

Yunyi Zhu received the master's degree from the National University of Singapore in Electrical Engineering under the guidance of Asst. Prof. Mike Zheng Shou in 2022. He is currently working as AI scientist with a startup company. His research interests include computer vision and image generation.

Pablo Arbeláez received the PhD degree with honors in applied mathematics from the Université Paris-Dauphine, in 2005. He was a senior research scientist with the Computer Vision Group, UC Berkeley from 2007 to 2014. He currently holds a faculty position with Universidad de los Andes in Colombia. His research interests are in computer vision, where he has worked on several problems, including perceptual grouping, object recognition, and the analysis of biomedical images.

David Crandall received the BS and MS degrees in computer science and engineering from the Pennsylvania State University, University Park, PA and the MS and PhD degrees in computer science from Cornell University, Ithaca, NY. He is luddy professor of Computer Science with Indiana University, where he is also director of the Luddy Artificial Intelligence Center. He has received an NSF CAREER Award, two Google Faculty Research Awards, an IU Trustees Teaching Award, a Grant Thornton Fellowship, and several best paper awards and nominations. Currently he is serving as an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence* and as a program chair of CVPR 2024 and ICDL 2024.

Dima Damen is a professor of computer vision with the University of Bristol and senior research scientist with Google DeepMind. She is currently an EPSRC fellow (2020-2025), focusing her research interests in the automatic understanding of object interactions, actions and activities using wearable visual (and depth) sensors. She is best known for her leading works in Egocentric Vision including the leading EPIC-KITCHENS dataset, and has also contributed to novel research questions including mono-to-3D, video object segmentation, video domain adaptation, skill/expertise determination from video sequences, dual-domain and dual-time learning as well as multi-modal fusion using vision, audio and language.

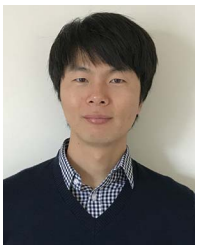
Giovanni Maria Farinella is full professor with the Department of Mathematics and Computer Science, University of Catania, Italy. His research interests lie in the fields of computer vision, pattern recognition and machine learning. He is associate editor of the international journals *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Pattern Recognition*, *International Journal of Computer Vision*. He founded (in 2006) and currently directs the International Computer Vision Summer School. He was awarded the PAMI Mark Everingham Prize 2017.

Christian Fuegen joined Meta as research scientist in 2013 due to the acquisition of Mobile Technologies. At Mobile Technologies, he was one of the core developers of "Jibbigo", an on-device speech-to-speech translator for mobile devices, where he worked from 2007 until 2013, first as research scientist and later as director of research. At Meta, his team particularly focuses on understanding complex (egocentric) acoustic scenes across multiple speakers and languages using audio and other modalities with the goal of advancing voice interfaces, including transcription, captioning, and content understanding for RayBan Stories, Oculus VR headsets, augmented reality, the Metaverse, and video understanding.

Bernard Ghanem received the BE degree from AUB in 2005 and the MS/PhD degrees from UIUC in 2010. He is a professor of CS/ECE at KAUST. His research interests lie in computer vision and machine learning, including video understanding, 3D recognition, and deep learning foundations. His work received several awards/honors, including a Abdul-Hameed-Shoman Arab Researcher Award for Big Data and Machine Learning (2020) and a Google Faculty Research Award (2015) [1st in MENA for Machine Perception]. He coauthored more than 175 papers. He has served as associate editor for *IEEE Transactions on Pattern Analysis and Machine Intelligence* and area chair (AC) for the main computer vision and machine learning conferences.

Vamsi Krishna Ithapu received the PhD degree in computer sciences from the University of Wisconsin-Madison. He is Machine Learning and Computer Vision Research Science manager and research lead with Reality Labs Research, Meta. His work is on the bridge of artificial intelligence (AI) systems and augmented/virtual reality (AR/VR). His group at develops and builds egocentric Audio-driven and Multi-Sensory Experiences by bringing together computational tools from AI, AR/VR, and human perception.

C. V. Jawahar is a professor with IIIT-H, Hyderabad. He leads the research group focusing on computer vision, machine learning, and multimedia systems. He has been actively involved in research questions that converge smart mobility, Indian language computing and multi-modal perception. He is an elected fellow of the Indian National Academy of Engineers (INAE) and the International Association of Pattern Recognition(IAPR). He was awarded the ACM India Outstanding Contribution to Computing Education (OCCE) 2021.



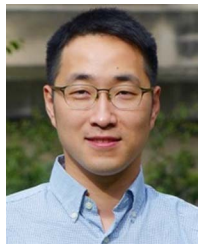
Hanbyul Joo received the BS and MS degree from KAIST and the PhD degree from the Robotics Institute, Carnegie Mellon University. He is an assistant professor with the Department of Computer Science and Engineering, Seoul National University (SNU). Before joining SNU, he was a research scientist with Facebook AI Research (FAIR), Menlo Park.

Kris Kitani received the BS degree from the University of Southern California and the MS and PhD degrees from the University of Tokyo. He is an associate research professor with the Robotics Institute, Carnegie Mellon University, and a research scientist with Meta FAIR. His research projects span the areas of computer vision, machine learning, and human computer interaction.

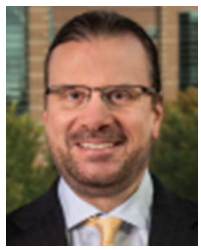
Haizhou Li (Fellow, IEEE) is the executive dean, and X.Q. Deng presidential chair professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also an adjunct professor with the National University of Singapore, Singapore. He was the editor-in-chief of *IEEE/ACM Transactions on Audio, Speech and Language Processing* (2015-2018), vice president of IEEE Signal Processing Society (2024-2026), president of International Speech Communication Association (2015-2017). He is an ISCA fellow, and fellow of Academy of Engineering Singapore. His research interests include speech information processing, natural language processing, and neuromorphic computing.

Richard Newcombe is VP of Research Science with Meta Reality Labs leading the Surreal team in Reality Labs Research. The Surreal team is creating a new generation of Machine Perception and egocentric AI technologies that combines novel always-on wearable sensing and compute with efficient algorithms for device location, 3D scene understanding and user state-estimation. The surreal team pioneered a new generation of glasses devices called project Aria that provides a new generation of data for ego-centric multimodal AI research.

Aude Oliva received the MS and PhD degree in cognitive science from the Institute National Polytechnique de Grenoble, France. She is a senior research scientist with the MIT Computer Science and Artificial Intelligence Laboratory where she heads the Computational Perception and Cognition group. She is the MIT director with the MIT-IBM Watson AI Lab. Her research is cross-disciplinary, spanning human perception, computer vision and neuroscience, and focuses on research questions at the intersection of all three domains. She has received an NSF Career Award in computational neuroscience, a Guggenheim fellowship in computer science and a Vannevar Bush Faculty Fellowship in cognitive neuroscience.



Hyun Soo Park received the PhD degree from Carnegie Mellon University. He is an associate professor with the Department of Computer Science and Engineering, the University of Minnesota. He is interested in modeling human and animal behaviors. Prior to joining the UMN, he was a postdoctoral fellow with the GRASP Lab, the University of Pennsylvania. He received NSF CAREER Award (2019) and CVPR 2021 Best Paper Honorable Mention.



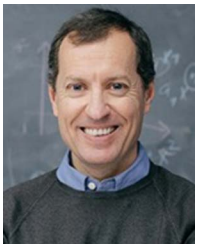
James M. Rehg received the PhD degree from CMU and taught with Georgia Tech from 2001-2022. He is a founder professor of computer science and industrial and enterprise systems engineering with UIUC, where he directs the Health Care Engineering Systems Center. He has received an NSF Career Award and a Raytheon Faculty Fellowship, and he and his students have received numerous best paper awards. He served as program co-chair for CVPR 2017 and general co-chair for CVPR 2009. His research interests include computer vision, machine learning, and mobile and computational health.

Yoichi Sato received the BS degree from the University of Tokyo in 1990 and the MS and PhD degrees in robotics from the School of Computer Science, Carnegie Mellon University in 1993 and 1997, respoectively. He is a professor with the Institute of Industrial Science, the University of Tokyo. His research interests include first-person vision, gaze sensing and analysis, and physics-based vision. He served/is serving in several conference organization and journal editorial roles including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, *Computer Vision and Image Understanding*, CVPR 2023 general co-chair, ICCV 2021 program co-chair, ACCV 2018 general co-chair, ACCV 2016 program co-chair and ECCV 2012 program co-chair.

Jianbo Shi received the PhD degree in computer science from the University of California at Berkeley in 1998 under Jitendra Malik for his thesis on the Normalized Cuts image segmentation algorithm. He is a professor with the GRASP laboratory, UPenn. He joined The Robotics Institute, Carnegie Mellon University in 1999 as a research faculty. In 2003, he joined the Department of Computer & Information Science, the University of Pennsylvania, where he is currently a professor. His current research focuses on human behavior analysis and image recognition-segmentation.

Mike Zheng Shou received the PhD degree from Columbia University. He is a tenure-track assistant professor with the National University of Singapore. He received the best paper finalist at CVPR'22, the best student paper nomination at CVPR'17. His team won the 1st place in the international challenges including ActivityNet 2017, EPIC-Kitchens 2022, Ego4D 2022 & 2023. He regularly serves as area chair for top-tier artificial intelligence conferences including CVPR, ECCV, ICCV, and ACM MM. He is a fellow of National Research Foundation Singapore.

Antonio Torralba received the graduation degree in telecommunications engineering from Telecom BCN, Spain, in 1994 and the PhD degree in signal, image, and speech processing from the Institut National Polytechnique de Grenoble, France, in 2000. He is the Delta electronics professor and head of the AI+D faculty with the Department of EECS, MIT. He received the 2010 J. K. Aggarwal Prize, the 2020 PAMI Mark Everingham Prize, the Inaugural Thomas Huang Memorial Prize by the PAMITC in 2021. In 2022, he was invested Honoris Causa doctor by the Universitat Politécnica de Catalunya-BarcelonaTech (UPC). He is a AAAI fellow.



Lorenzo Torresani received the Laurea degree in computer science with summa cum laude honors from the University of Milan (Italy) in 1996, and the MS and PhD degrees in computer science from Stanford University in 2001 and 2005, respectively. He is a research director with Facebook AI Research (FAIR), Meta. From 2009 to 2021, he was on the faculty of the Computer Science Department with Dartmouth College, where he received tenure in 2014 and was promoted to full professor in 2020. He is the recipient of several awards, including a CVPR best student paper prize, an NSF CAREER Award, a Google Faculty Research Award, Facebook Faculty Awards, and a Fulbright U.S. Scholar Award.

Mingfei Yan received the master's of computer engineering from Carnegie Mellon University. She leads the product development team in Meta's Reality Labs Research Division, focusing on envisioning and developing the next-gen computer vision and machine perception technologies. She has been supporting Project Aria's development since its inception. Prior to joining Meta, she worked with Microsoft Hololens and Microsoft Azure as product manager.

Jitendra Malik (Fellow, IEEE) is Arthur J. Chick professor of EECS with UC Berkeley, and research scientist director with FAIR, Meta Inc. Over his career, he has advised more than 80 PhDs and postdoctoral fellows. He received the 2013 PAMI Distinguished Researcher Award in Computer Vision, and the 2019 IEEE Computer Society's Computer Pioneer Award for "leading role in developing Computer Vision into a thriving discipline through pioneering research, leadership, and mentorship". He is fellow of ACM and the American Academy of Arts and Sciences, and a member of the National Academy of Sciences & the National Academy of Engineering.