HD-EPIC: A Highly-Detailed Egocentric Video Dataset

Toby Perrett*♣ Ahmad Darkhalil*♣ Saptarshi Sinha*♣ Omar Emara*♣ Sam Pollard*♣
Kranti Kumar Parida*♣ Kaiting Liu,*♠ Prajwal Gatti*♣ Siddhant Bansal*♣ Kevin Flanagan*♣
Jacob Chalk*♣ Zhifan Zhu*♣ Rhodri Guerrier*♣ Fahd Abdelazim*♣ Bin Zhu♦
Davide Moltisanti♥ Michael Wray♣ Hazel Doughty♠ Dima Damen♣
♣Uni. of Bristol ♠Leiden Uni. ♦Singapore Management Uni. ♥Uni. of Bath *: Equal Contribution

http://hd-epic.github.io

Abstract

We present a validation dataset of newly-collected kitchen-based egocentric videos, manually annotated with highly detailed and interconnected ground-truth labels covering: recipe steps, fine-grained actions, ingredients with nutritional values, moving objects, and audio annotations. Importantly, all annotations are grounded in 3D through digital twinning of the scene, fixtures, object locations, and primed with gaze. Footage is collected from unscripted recordings in diverse home environments, making HD-EPIC the first dataset collected in-the-wild but with detailed annotations matching those in controlled lab environments.

We show the potential of our highly-detailed annotations through a challenging VQA benchmark of 26K questions assessing the capability to recognise recipes, ingredients, nutrition, fine-grained actions, 3D perception, object motion, and gaze direction. The powerful long-context Gemini Pro only achieves 37.6% on this benchmark, showcasing its difficulty and highlighting shortcomings in current VLMs. We additionally assess action recognition, sound recognition, and long-term video-object segmentation on HD-EPIC.

HD-EPIC is 41 hours of video in 9 kitchens with digital twins of 413 kitchen fixtures, capturing 69 recipes, 59K fine-grained actions, 51K audio events, 20K object movements and 37K object masks lifted to 3D. On average, we have 263 annotations per minute of our unscripted videos.

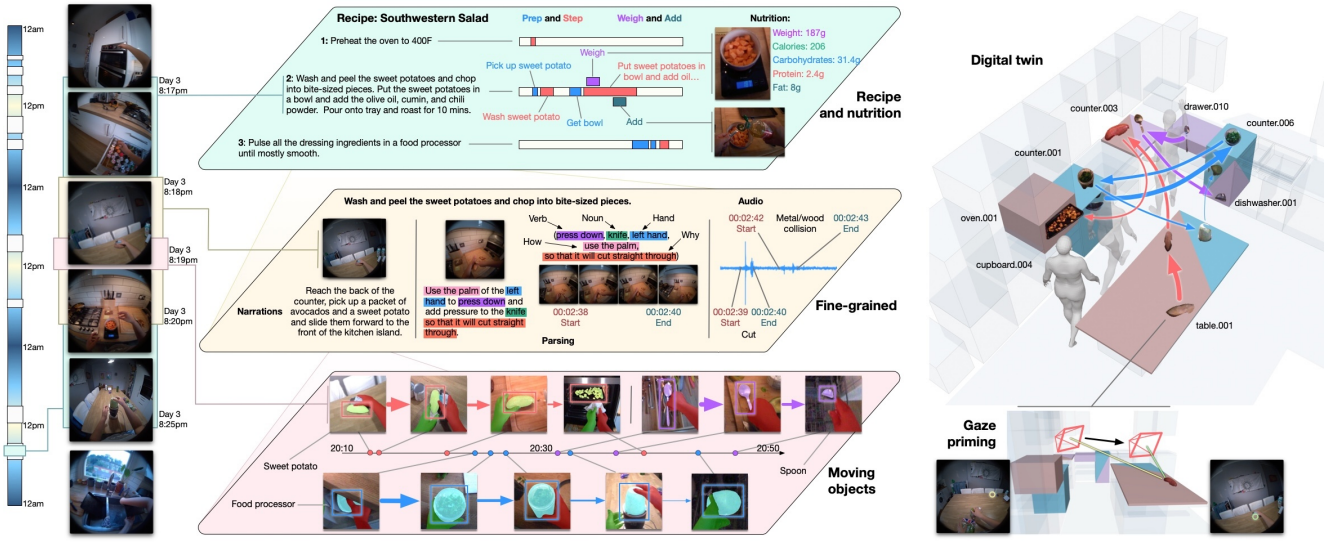

Figure 1. **Annotation Highlights.** We capture multi-day recordings of unscripted activities. **Centre-Top**: Recipes are recorded with steps and their preparation temporally annotated, along with ingredient addition. Ingredients are weighed and nutrition recorded. **Centre-Middle**: Dense fine-grained narrations detailing what, how, and why are parsed and clustered. Audio events are also annotated. **Centre-Bottom**: Object movements are temporally annotated with bounding boxes and hands and object masks. **Right-Top**: All annotations are temporally grounded in a 3D digital twin. We show trajectories of 3 (masked) objects: Sweet potato, Food processor and Spoon, highlighting relevant kitchen fixtures. **Right-Bottom**: Gaze captures when objects are primed (*i.e.* looked at) before being taken/placed.

1. Introduction

Detailed understanding of videos, from the brief fine-grained action to the overarching hour-long activity, is effortless for humans but currently out of reach for both foundational and specialised models. Egocentric videos, in particular, introduce additional challenges to general video understanding, including significant camera motion, subtle action motion, objects occluded during manipulations and frequently going out of view. Understanding such videos requires disentangling the combined signals of head motion, hand interactions and a global understanding of the dynamic scene. This makes ego videos a great testbed for a comprehensive evaluation of video perception models.

Egocentric vision has recently been fuelled by an influx of datasets [18, 28, 29, 80]. While large-scale, making them ideal for training, these datasets are sparsely annotated, particularly for tasks which link various parts of the long video, or those requiring 3D grounding. In contrast, richly annotated datasets tend to be synthetic or collected in controlled settings [7, 50, 68] which limits their realism. We bridge this gap by presenting the **most densely annotated dataset of unscripted recordings**, ideal for comprehensive **validation** of video-only and video-language models.

We collect new videos, allowing us to capture additional meta-data and to ensure these videos have not already been used to train existing models. Following EPIC-Kitchens [17], participants collect all kitchen activities for three days. We thus term our dataset Highly-Detailed EPIC (HD-EPIC). Fig. 1 provides an overview of the multi-tiered annotations, several of which are novel:

- ⋆ Recipe steps are temporally annotated, and linked to annotations of all preparatory actions that relate to the step.
- ⋆ Ingredients are weighed in videos and labelled with nutrition. We track dish nutrition as ingredients are added.
- ⋆ Each action has a dense description capturing the what, how, and why of actions along the start and end time.
- ⋆ For each kitchen, we curate a digital twin with labelled fixtures. These are associated with actions (*e.g.* open/close) and the taking/placing of objects.
- ⋆ All moved objects are tracked, with manual masks lifted to 3D bounding boxes.
- ⋆ We associate gaze with object movements, labelling when objects are spotted before take/place actions.

With these dense annotations, we design a challenging Visual Question Answering (VQA) benchmark of 26K questions. We purposefully do not use LLMs to generate negatives, instead using similar annotations. We highlight a few novel question types:

- ⋆ Recipe nutrition: we question the change in the recipe nutrition as one or more ingredients are added.
- ⋆ Multi-video: we question recipes prepared across recordings, with a VQA that spans multiple long videos.
- ⋆ Object itinerary: we question multi-hop object movements over a long video, relative to kitchen fixtures.
- ⋆ Fixture interactions: we question how many times a particular cupboard/drawer is opened/closed.
- ⋆ Action how/why: we question how/why an action was carried out, using participant-narrated manners/reasons.
- ⋆ Anticipation with gaze: With gaze priming, we query next-object movement, offering evidenced anticipation.

Additionally, we report results on action recognition, sound recognition, and long-term video object segmentation.

This paper thus contributes: (i) 41 hours of multi-day unscripted egocentric recordings, (ii) highly-detailed annotations including novel labels (*e.g.* ingredient nutrition, digital twin, gaze prime) and (iii) a challenging VQA benchmark including novel Qs (*e.g.* object itinerary, recipe nutrition changes) along with 3 standard video benchmarks.

2. Related Work

With the rise of foundation models [5, 6, 11, 13, 21, 59, 70, 79, 81, 88], there has been a recent influx of benchmarks [8, 10, 15, 23, 25, 35, 42–44, 49, 58, 78, 92] aiming to test video understanding abilities. These benchmarks evaluate diverse capabilities *e.g.* physics [58], counting [23], temporal reasoning [8, 10] and long video [23, 25, 49].

A few benchmarks test embodied or egocentric understanding. [28] released a Natural Language Queries (NLQ) benchmark (19.2K queries) centred around episodic memory of objects. [48] collects 1.6K human-made questions and answers on topics such as relative object locations, episodic memory, and spatial reasoning. However, it uses views from the HM3D [63] and ScanNet [16] datasets, so these questions are based on passive views of a static environment. [49, 86] auto-generate 5K and 7K questions based on Ego4D narrations. Whilst this approach is efficient, it is limited to these short narrations. [8] collects its own annotations for videos from several datasets, including Ego4D. Their benchmark is solely focused on temporal questions related to ordering, counting, causality and direction.

To evaluate a wider range of capabilities, a wider range of annotations are required. Of particular note are 3D grounding annotations. Ego4D [28] contains some environment scans and static 3D object locations. With SLAM-equipped devices [91] builds a benchmark for 3D object tracking; [50] contains an office and living room digital twin; and [29] contains ego- and exo- views of expert tasks.

In contrast to these works which focus on only a few annotation types, we collect the most comprehensive set of annotations in one dataset, including highly detailed narrations, object and hand segmentations, and a comprehensive 3D digital twin of the scene and objects, all from unscripted egocentric footage in participants' homes.

3. Data Collection

Recruitment and Equipment. Each participant engaged in a long commitment (~50 hours) involving data record-

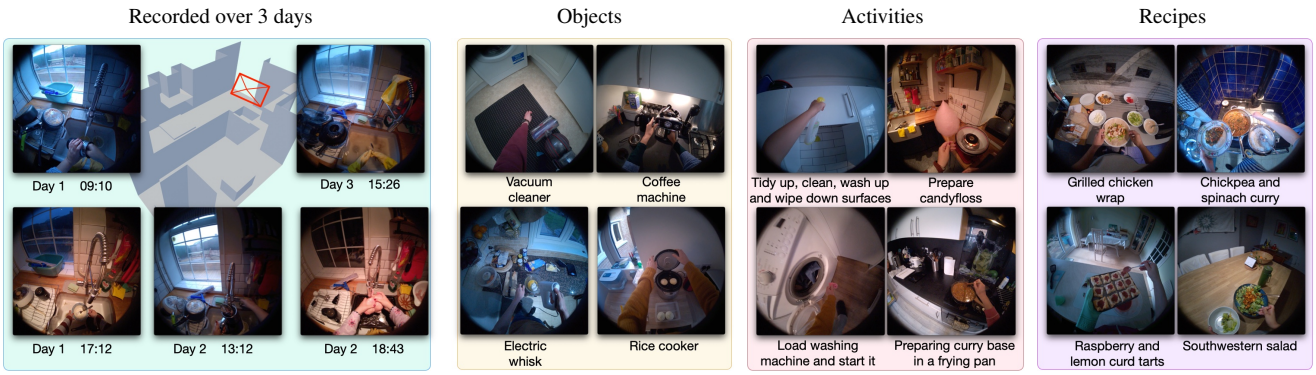

Figure 2. Diversity in HD-EPIC, which is filmed over 3 days in-the-wild, resulting in many objects, activities and recipes.

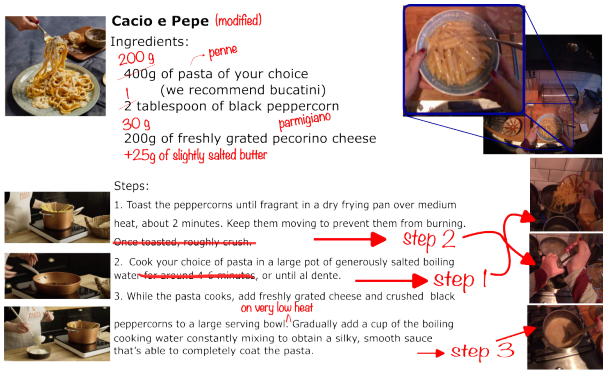

Figure 3. Recipe modification in ingredients and steps.

ing and providing detailed narrations, recipes and nutrition information. Data was collected with Project Aria glasses [72]—a multi-sensor platform with 3 forward cameras (1 RGB and 2 SLAM), 7 microphones and inward cameras for gaze estimation. We collected 30 FPS RGB videos at 1408×1408 resolution, 60 FPS eye tracking and 30 FPS SLAM. We supplied participants with multiple devices including scales for nutritional tracking (see Fig. A1).

Instructions and Collected Data. Participants recorded all their daily kitchen activities for at least 3 consecutive days. All 9 participants were asked to wear the glasses each time they walked into their kitchen, pressing record upon entering, and stopping the recording when they left the kitchen. Participants recorded for 3.5 to 7.2 hours (avg. 4.6). Overall, we collected 156 videos, with an average length of 15.9 ($\pm$ 14.5) minutes totalling 41.3 hours (4.46M frames). Fig. 2 shows the diversity in the collected data.

Following data collection, participants provided the recipes they freely prepared, citing the source (*e.g.* website) and any modifications (see Fig. 3). We collected a total of 69 recipes covering various cuisines. On average, recipes contained 6.6 steps, 8.1 ingredients, and took 4 hours 48 minutes across 2.1 videos from preparation to serving. Our longest recipe took 2 days and 6 hours to complete.

To track nutrition of recipes, participants weighed and manually logged ingredients with MyFitnessPal [3], giving us detailed nutrition information and adding an additional dimension to the dataset. In total, participants used 558 ingredients including ingredients high in protein, *e.g.* tuna and kidney beans; carbohydrates, *e.g.* dates and flour; and fat *e.g.* sour cream and pine nuts. Participants prepared both high calorie dishes *e.g.* Lazy Cake (4.8K calories) and low calorie dishes *e.g.* Crispy Cucumber Salad (274 calories).

Narrations. We follow prior datasets [17, 18, 28], asking participants to watch their recordings and narrate with a web-based narrator tool [28]. We expand on this by asking participants to describe *what* they are doing, along with *how* and *why*. This results in a rich set of narrations that are denser, and more detailed than previous datasets (*e.g.* 3.8× more words/min than Ego4D). See stats in Supp. B.

Post-Processing—Multi-Video Slam and Gaze. We use Aria MPS [1] to process videos obtaining singular multi-day point clouds per kitchen; 1kHz 6DoF camera trajectories; and eye gaze direction. We post-process VRS files, converting videos to mp4, removing the gaze camera input for anonymity. Further details are in Supp. B.

4. Annotation Pipeline

We collect extensive multi-tiered annotations to achieve the level of detail that distinguishes HD-EPIC from other video understanding datasets. Here we detail our pipeline.

4.1. Annotating Recipe Steps and Ingredients

Our videos are distinct from short recipe videos found online, which are typically trimmed to only crucial steps, and often edited further or sped up. Videos in HD-EPIC include a wider range of recipe-relevant activities, such as fetching or prepping ingredients. To comprehensively annotate these videos, we introduce prep and step pairs.

The *prep* of a corresponding *step* is defined as all essential actions the participant takes to get ready to execute a given step. For example, the prep of the step 'chop tomato', includes retrieving the tomato from storage, washing it, and gathering the knife and chopping board. However, if the step is 'Add chopped onions and stir', then the chopping of onions is part of the prep for that step. This introduces a more fine-grained understanding of all steps, unexplored in prior datasets [29, 39, 73]. Fig. 4 shows sample prep-step

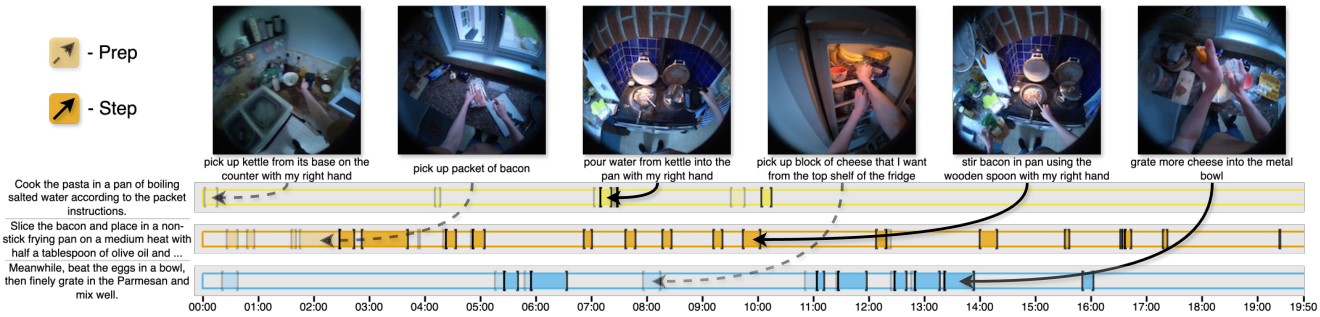

Figure 4. For the 'Carbonara' recipe, we visualise the *prep* and *step* time segments for three consecutive steps (left), along with sample frames with corresponding action narrations (top). The interleaving of different *preps*/steps is evident in the annotations.

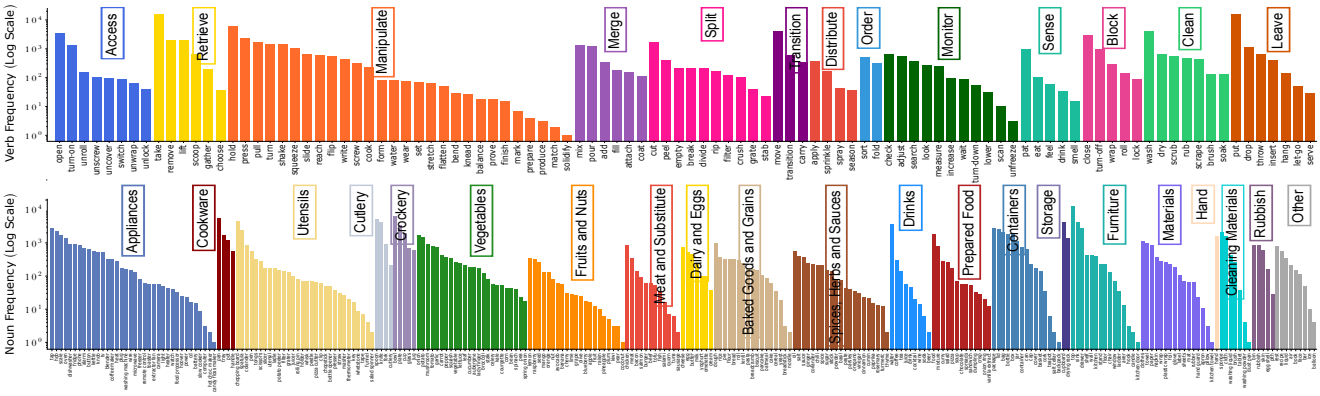

Figure 5. Frequency of verb clusters (top) and noun clusters (bottom) in narrated sentences by category, shown on a logarithmic scale.

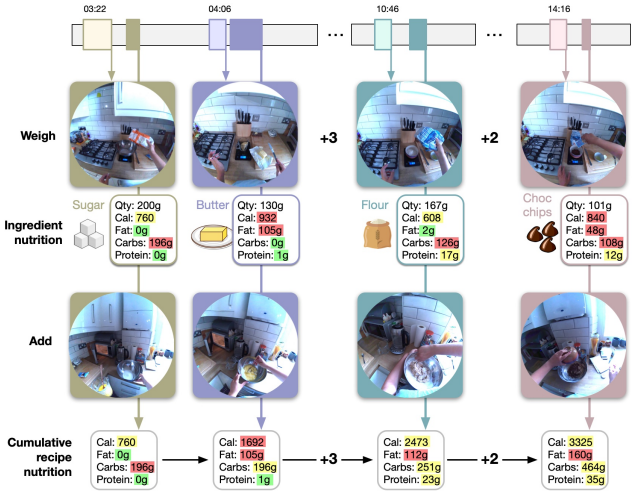

Figure 6. Nutrition is monitored throughout recipes as ingredients are incorporated into dishes. Here we show Banana Bread Chocolate Chip Cookies. We annotate when ingredients are weighed, document their nutrition, and locate their adding time, giving us overall dish nutrition at each stage.

annotations for 3 steps. Nearly all steps (93.1%) have paired prep annotations. Typically, prep is shorter than a step: avg. prep is 54.5s (±95.3s), avg. step is 78.2s (±100.7s).

We also annotate *weighing* and *adding* temporal segments which enables monitoring the nutrition of the full dish as ingredients are incorporated (see Fig. 6). In total, we annotate 283 in-view weighing sequences (avg. 18.9s) and 501 adding sequences (avg. 31.6s), excluding spices. Details of the annotation process are in Supp. C.

4.2. Fine-Grained Actions

Transcription. We automatically transcribe and manually check and correct all audio narrations provided by participants, to obtain detailed action descriptions.

Action Boundaries. For all narrations, we label precise start and end times. In total, we obtain segments for 59,454 actions, with a mean duration of 2.0s (±3.4s).

Parsing. We parse verbs, nouns and hands from open vocabulary narrations so they can be used for closed vocabulary tasks, such as action recognition. We also extract how and why clauses from 16,004 and 11,540 narrations, respectively. For example, "Turn the salt container clockwise by pushing it with my left hand so that the lid is aligned with the container opening."

Clustering. Fig. 5 shows the distribution of clusters (*i.e.* classes) across all videos in HD-EPIC, along with hierarchical clusters [18]. As with prior datasets [18, 28], our highly diverse actions and objects are long-tailed.

Sound Annotations. We follow [31] to collect audio annotations. These capture start-end times of audio events along with a class name (e.g. 'click', 'rustle', 'metal-plastic collision', 'running water'). Overall, we have 50,968 audio annotations from 44 classes.

Full details of transcription, boundary labelling, parsing, clustering and sound annotations are in Supp. C.

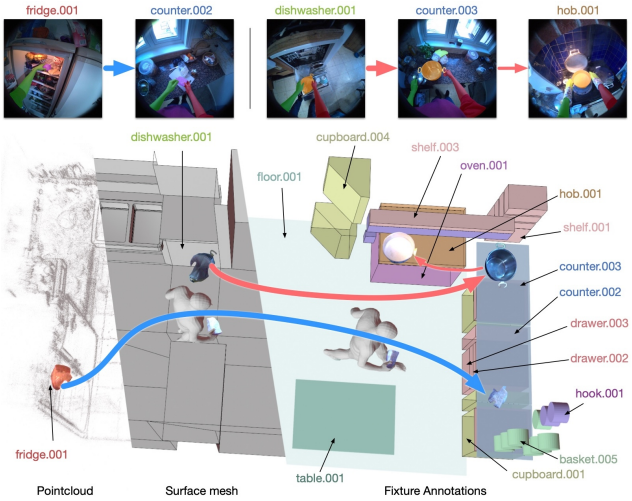

Figure 7. Digital Twin: from point cloud (left), to surfaces (middle) and labelled fixtures (right). We show two moved objects (masks on top) at fixtures: cheese and pan. Body poses from [87].

4.3. Digital Twins: Scene & Object Movements

Scene. We create digital copies of participants' kitchens by reconstructing the surfaces and manually curating every fixture (*e.g.* cupboard, drawer), storage space (*e.g.* shelves, hooks) and large appliance (*e.g.* fridge, microwave). This is distinct from digital twins that rely on known environments with replicas. Our digital twin is created in Blender [2] on top of the multi-video SLAM point clouds from recordings.

Each kitchen contains an average of 45.9 labelled fixtures (min 31, max 62), including 14.2 counters/surfaces, 12.2 cupboards, 7.8 drawers and 5.2 appliances (sample in Fig. 7). We refer to these annotations as Fixtures F.

We then associate narrations which describe scene interactions with F. We find actions where a noun indicates a fixture, *e.g.* "open drawer", identify the exact "drawer" in the digital twin (*e.g.* drawer.001) and update its state. Following studies showing humans fixate up to 1 second before interacting [40], we take the fixture $f \in F$ with the highest cumulative gaze intersection for the 1s before the narration.

Hand Masks. We annotate a handful of frames per video for both hands. Frames are selected to cover various actions and kitchen locations. We use these to automatically segment, and manually correct a selected subset. In total, our dataset contains 7.7M hand masks: 3.9M right and 3.8M left of which 11K are manually annotated (details in Supp. C).

Moving Objects in 2D. To generate 3D object movement annotations, we first annotate when objects move. Annotators label a temporal segment each time an object is moved until it is set, along with 2D object bounding boxes at the onset and end of motion. For example, if a person moves a cup from a countertop to the sink, one bounding box captures the cup on the countertop and another when in the sink. Tracks are annotated even for slight shifts/pushes, and thus offer full annotations of all object movements.

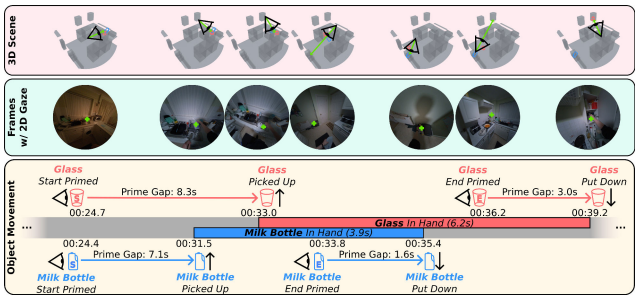

Figure 8. **Priming Object Interaction Through Gaze**. Top: Camera position with projected eye-gaze and object positions in 3D. Middle: 2D gaze location. Bottom: Timeline for priming object movement *e.g.* the glass is primed 8.3s before taking.

Overall, we collected 19.9K object movement tracks and 36.9K bounding boxes. We label an average of 9.2 objects taken and 9.0 objects placed per minute. On average, tracks are 9.0s long, the longest is 461.5s.

Object Masks. Despite progress in segmentation [36, 64] and available annotations [19, 29], models perform poorly in egocentric video, particularly under occlusions. We obtain pixel-level segmentations from each bounding box by initialising with iterative SAM2 [64] then manually correcting. Annotators corrected 74% of masks; the IoU between SAM2 and the manual masks is 0.82.

Masks to 3D. We lift object masks to 3D using dense depth estimates and 2D-to-3D sparse correspondences provided by MPS. Given metric depth from [85], we identify S, the set of pixels within or around the object with 3D correspondences. We then find the linear transformation coefficients: $\alpha, \beta = \underset{\alpha,\beta}{\text{argmin}} \|(\alpha \hat{D}_S + \beta) - D_S\|^2$. where $\hat{D}_S$ are estimated depth values and D_S are existing depth values, followed by RANSAC to remove outliers.

3D Object Motion. Objects move 61.4cm ($\pm$84.5cm) on average, 27.6% move $\leq$10cm, while 7.6% move $\geq$2m.

Object-Scene Interactions. With the 3D object locations, we associate locations with the closest fixture $f \in F$, subject to fixture-specific heuristics (*e.g.* objects must be within a counter's x-y plane). We manually verify all assignments, correcting any errors. On average objects move between 1.8 different fixtures per video (see Supp. C for stats).

Priming Object Movement. The behaviour of gaze when picking up and placing objects is well-studied [32, 40]. We combine eye-gaze and 3D object locations, to find when an object is *primed*, *i.e.* the moment in time when the gaze attends to the object's location before picking it up (*pick-up priming*) or when the gaze attends to the future location of an object before it's put down (*put-down priming*).

We calculate the *priming time* for all objects, excluding those taking or placed off screen. Additionally, at times, a person is already manipulating an object well before picking it up. We thus exclude objects with a pick up location

Dataset	Val&Test Hours	Action Segments	Unscripted	Free Setting	Recipe	Nutrition	Gaze	Audio Labels	Object	Hand	3D object over time	Labelled 3D environment	Camera pose	Fully annotated
HOI4D [46]	11.4	✓	✗	✗	✗	✗	✗	✗	Mask	Mask	✓	✓	✗	✓
Assembly101 [68]	66.8	✓	✗	✗	✗	✗	✗	✗	✗	3D pose	✗	✗	✗	✓
EPIC-KITCHENS-100 [18]	25.3	✓	✓	✓	✗	✗	✗	✓	Mask	Mask	✓	✗	✓	✗
Ego4D [28]	288.7	✓	✓	✓	✗	✗	✗	✗	B-Box	B-Box	✗	✗	✗	✗
HoloAssist [80]	49.8	✓	✗	✗	✗	✗	✓	✗	✗	3D pose	✗	✗	✓	✓
Aria Digital Twin [50]	8.1	✗	✗	✗	✗	✗	✓	✗	Mask	✗	✓	✓	✓	✓
Aria Everyday Activities [47]	7.3	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗
Aria Everyday Objects [74]	0.4	✗	✓	✓	✗	✗	✗	✗	B-Box	✗	✓	✗	✓	✓
Ego-Exo4D [29]	85.1	✗	✓	✓	✓	✗	✓	✗	Mask	3D pose	✗	✗	✓	✗
HD-EPIC	41.3	✓	✓	✓	✓	✓	✓	✓	Mask	Mask	✓	✓	✓	✓

Table 1. **Comparison of Egocentric Video Datasets** (see full table A1 in Supp.).

Location	Filtered (% Total)	Feasible (% Total)	Primed (% Feasible)	Not Primed (% Feasible)	Avg. Time (s)
Start	29.40	70.60	94.82	5.18	3.99(±2.94)
End	66.92	33.08	88.46	11.54	2.62(±2.05)

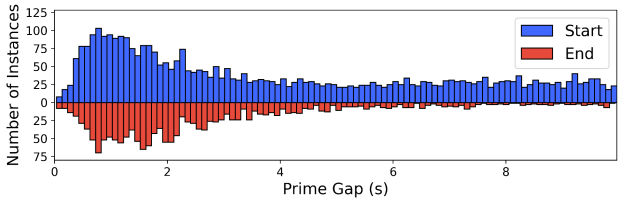

Figure 9. (Top) Priming Statistics for both start and end locations (Bottom) Histogram showing the difference in time when an object is primed before it is picked up (blue) or placed (red).

already close to the gaze 10s earlier. In Fig. 8 we show gaze priming for two objects: milk bottle and glass. The glass's end location, a cupboard, is primed 3s before the glass is put away. Fig. 9 displays priming statistics. Of those objects feasible for priming, 94.8% are primed, an average of 4.0s before being picked up, compared to 88.5% primed an average of 2.6s before being placed.

Long Term Object Tracking. We connect object movements and form longer trajectories, *i.e.* object itineraries, to capture sequences of an object's movement. Our efficient pipeline utilises our lifted 3D locations and allows annotating a 1-hour long video in minutes (details in Supp. C).

4.4. HD-EPIC vs Prior Egocentric Datasets

Tab. 1 compares HD-EPIC to other egocentric datasets (full table in Supp.). Compared to the largest dataset with labelled 3D environments (Aria Digital Twin [50]), HD-EPIC contains 5x more footage; has more annotations; and importantly was collected in an unscripted manner in the participants' homes. In particular, HD-EPIC is the first to annotate recipes, nutritional values, detailed action segments, gaze and audio labels on the same set of videos. With these diverse and dense annotations, HD-EPIC constitutes a true zero-shot benchmark for video understanding.

5. Benchmarks and Results

We show the potential of HD-EPIC as a validation dataset with benchmarks on general Video Question Answering (VQA) (Sec 5.1), action and sound recognition (Sec 5.2) and long-term video object segmentation (Sec 5.3).

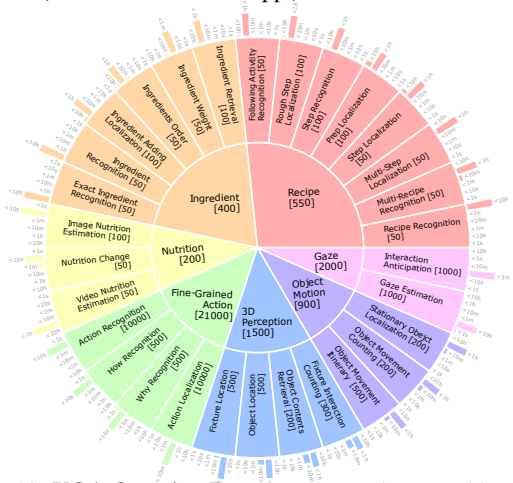

Figure 10. **VQA Question Prototypes**. We show our 30 question prototypes by category alongside the number of questions. Outer bars indicate the distribution over input lengths for each question.

5.1. HD-EPIC VQA Benchmark and Analysis

Benchmark Creation. We take the dense output of our annotation pipeline and construct a comprehensive VQA benchmark around 7 types of annotations:

1. Recipe. Questions on temporally localising, retrieving, or recognising recipes and their steps.
2. Ingredient. Questions on the ingredients used, their weight, their adding time and order.
3. Nutrition. Questions on nutrition of ingredients and nutritional changes as ingredients are added to recipes.
4. Fine-grained action. What, how, and why of actions and their temporal localisation.
5. 3D perception. Questions that require the understanding of relative positions of objects in the 3D scene.
6. Object motion. Questions on where, when and how many times objects are moved across long videos.
7. Gaze. Questions on estimating the fixation on large landmarks and anticipating future object interactions.

For each question type, we define prototypes to sample questions, correct answers, and strong negatives from our annotations. For example, Object Movement Counting asks "How many times did the object <bbox> seen at <time> move in the video?". This uses long videos, requiring multiple hops to be correctly answered. In contrast,

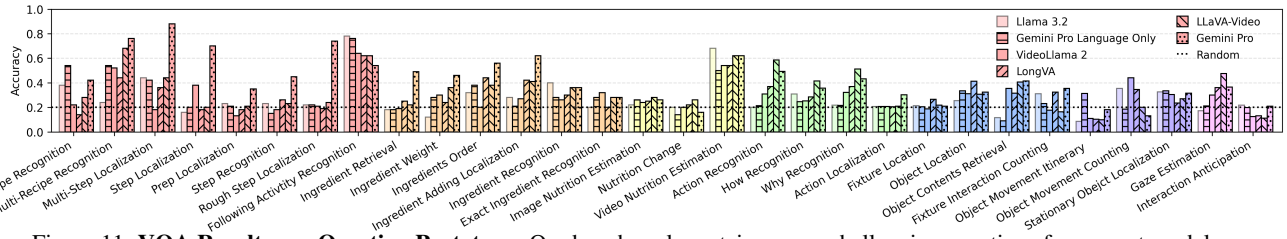

Figure 11. **VQA Results per Question Prototype**. Our benchmark contains many challenging questions for current models.

Model	Recipe	Ingredient	Nutrition	Action	3D	Motion	Gaze	Avg.
Blind - Language Only								
Llama 3.2	33.5	25.0	36.7	23.3	22.3	25.5	19.5	26.5
Gemini Pro	38.0	26.8	30.0	22.1	21.5	27.7	20.5	26.7
Video-Language								
VideoLlama 2	30.8	25.7	32.7	27.2	25.7	28.5	21.2	27.4
LongVA	29.6	30.8	33.7	30.7	32.9	22.7	24.5	29.3
LLaVA-Video	36.3	33.5	38.7	43.0	27.3	18.9	29.3	32.4
Gemini Pro	60.5	46.2	34.7	39.6	32.5	20.8	28.7	37.6
Sample Human Baseline	*96.7*	*96.7*	*85.0*	*92.5*	*93.8*	*92.7*	*75.0*	*90.3*

Table 2. **VQA Results per Category** (% Acc.). Our VQA benchmark cannot be solved blind or by external knowledge and is a challenge for state-of-the-art video VLM models.

How Recognition asks "What is the best description for how the person carried out the action <verb, noun>?" to test a model's ability to capture intricate details of actions.

Each question prototype is 5-way multiple choice. We generate hard negatives for prototypes by sampling within the dataset for difficult answers. For example, we take 4 different answers of how participants performed the same action. This ensures realistic negatives and challenging questions. In total, we have 30 prototypes, and generate 26,650 multiple-choice questions. This makes it one of the largest VQA video benchmarks, but keeps it tractable particularly to evaluate closed-source VLMs. Due to the density of our annotations, we estimate an upper bound of 100,000 possible unique questions with this set of prototypes.

Fig. 10 shows the distribution of questions per category alongside the distribution of input lengths which varies from single frames to 7+ hours. Details of each prototype's sampling can be found in Supp. D. A sample of our questions and answers can be seen in Fig. 13.

VLM Models. Due to the size and long-term nature of many question prototypes in our benchmark, we use 5 representative models as baselines (more details in Supp. D):

- Llama 3.2 90B [21]. We use this as a strong open-source (OS) text-only baseline, as LLMs can perform well on visual QA benchmarks *without any visual input* [82].
- VideoLlama 2 7B [13]. OS short context model.
- LongVA [89]. Longest context OS model.
- LLaVa-Video [90]. OS model trained also on ego data.
- Gemini Pro [75]. Closed source, longest context of any model, and state-of-the-art on long-video [25].

VQA Results Per Category and Per Prototype. Tab. 2 provides overall and per-category accuracy averaged over the prototype results shown in Fig. 11. Both language-only models only achieve 26.5% and 26.7%, only 6.7% above random. Open-source video VLMs (VideoLlama, LongVA, LlaVA-Video) perform similarly (27.4%, 29.3%, 32.4%) but have different strengths as shown in Fig. 11. For example, Llama better estimates nutrition, while the video is necessary to get above random performance on action recognition and gaze estimation. Gemini achieves the best performance, particularly for Recipe and Ingredient where external knowledge helps. However, the average performance (37.6%) and the gap to our sample human baseline (90.3%) shows the challenge posed by our VQA benchmark.

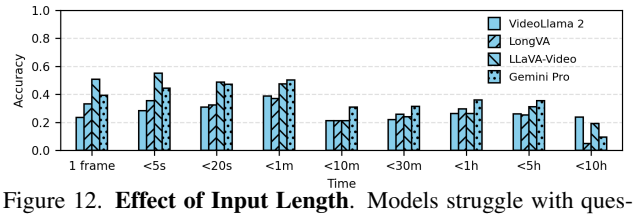

Figure 12. **Effect of Input Length**. Models struggle with questions of all video input lengths. s=second, m=minute, h=hour.

Model	Modality	Verb	Noun	Action	Unseen EPIC-100 Action
EPIC-KITCHENS-100 SOTA					
TIM [9]	A+V	77.1	67.2	57.5	44.6
HD-EPIC					
Chance	-	10.9	1.8	0.0	-
SlowFast [24]	V	29.2	10.6	5.3	29.0
Omnivore [26]	V	19.5	17.1	8.7	28.7
MotionFormer-HR [51]	V	35.7	20.0	10.2	32.2
VideoMAE-L [76]	V	47.5	29.4	17.9	29.3
TIM [9]	A+V	**51.3**	36.1	23.4	**44.6**
TIM [9]	V	51.2	**36.5**	**23.9**	44.4

Table 3. **Action Recognition Benchmark** (% Acc.). HD-EPIC provides a significant challenge for state-of-the-art models.

Video Length. Fig. 12 shows models struggle with all video lengths but are worst with inputs ≥1 minute.

Common Failures. Fig. 13 shows qualitative results. In Recipe, models struggle when steps have common objects or actions. In Ingredient, models guess weights (readable from the scale by humans) poorly, also causing errors in Nutrition. Fine-grained action is hard when answers share nouns. In Gaze, models just select recently moved objects. Confusion in 3D and Object motion occurs with directions (right/left) and fixtures (counters/drawers).

5.2. Recognition Benchmarks

Action Recognition. We assess 5 action recognition methods [9, 24, 26, 51, 76], using publicly available checkpoints fine-tuned on EPIC-KITCHENS-100. Results are shown in Tab. 3. For context we show the results from EPIC-KITCHENS-100 (top row) and on the unseen kitchens subset of EPIC-KITCHENS-100 (last col.). Best performance

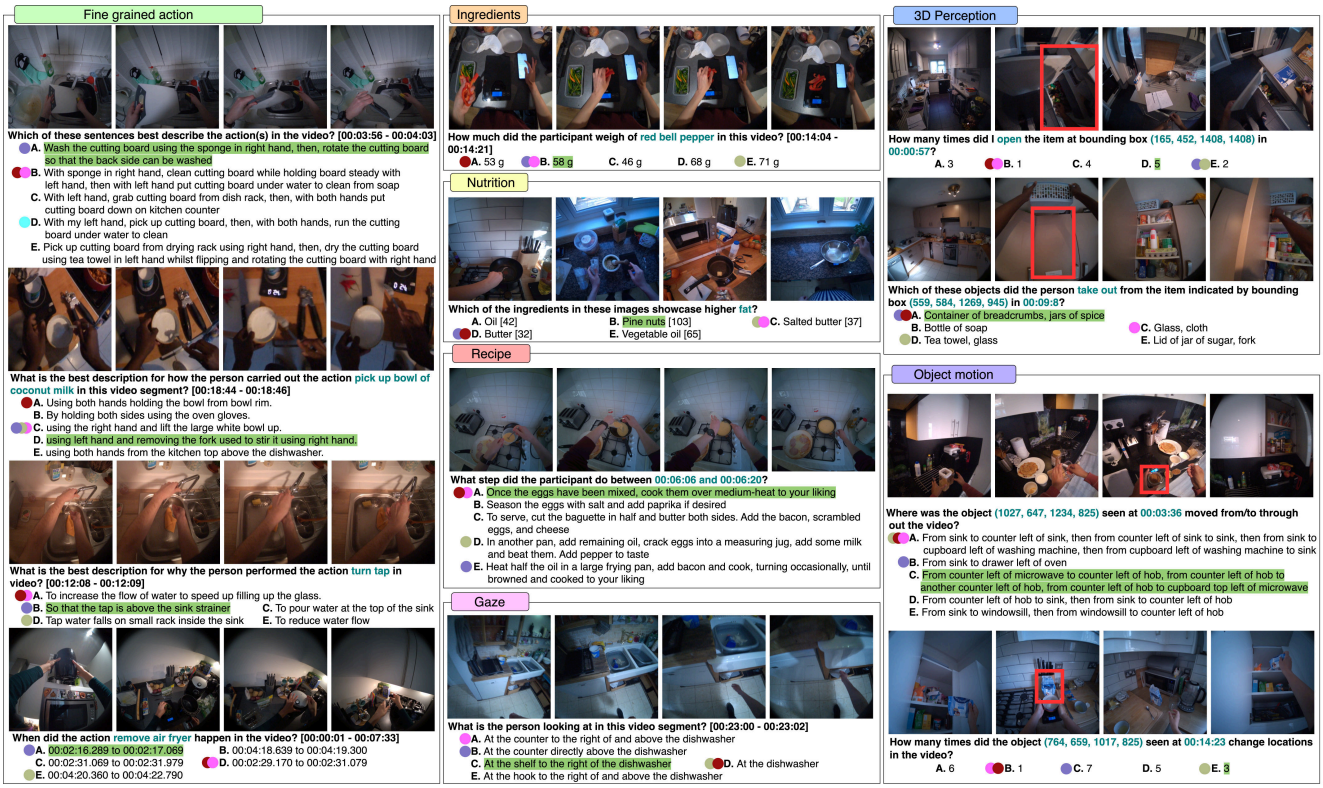

Figure 13. **VQA Qualitative Results**. We mark GT answers with a green background, and predictions from different models, *i.e.*, LLaMA 3.2, VideoLLaMA 2, LongVA, Gemini Pro with coloured dots. Note: Under Nutrition, [fat] values are not provided to the model.

Model	Modality	Top-1	Top-5	mCA	mAP	mAUC
EPIC-Sounds SOTA						
TIM [9]	A+V	58.3	86.0	25.8	30.6	0.879
HD-EPIC						
Chance	-	6.9	29.4	2.2	2.3	0.500
SSAST [27]	A	25.1	59.8	10.8	13.5	0.748
TIM [9]	A	26.9	56.9	12.4	11.4	0.689
ASF [34]	A	27.9	**64.0**	11.9	14.0	0.741
TIM [9]	A+V	**31.9**	61.0	**14.4**	**15.7**	**0.765**

Table 4. **Sound Recognition Benchmark**. Current models struggle on HD-EPIC compared to the EPIC-Sounds state-of-the-art.

on HD-EPIC is only 51% for verbs, 37% for nouns and 24% for actions leaving plenty of room for improvement.

Sound Recognition. We evaluate 3 audio models [9, 27, 34], all trained on EPIC-Sounds. Tab. 4 shows a large gap in performance comparing HD-EPIC to EPIC-Sounds for SSAST (-28.4), ASF (-25.9) and TIM (-26.4). This shows audio is not sufficiently robust to new scenes or devices.

5.3. Long-Term VOS Benchmark

We construct a long-term video object segmentation benchmark using our segmentations and track associations (Sec. 4.3). Our benchmark has 1000 sequences, each with 1-5 objects and 2 hand masks (see Supp. D for details). While we have a lot more tracks, we keep it comparable to current benchmarks in size. We evaluate two models [12, 64] with a naive baseline where object masks are kept static. Fig. 14 shows the results. SAM2 [64] surpasses Cutie [12] for hands, but does worse on objects. Overall, objects have added challenge in diversity in perspective, lighting, location and occlusion.

Model	Total			Hands			Objects		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Static	8.0	10.3	9.2	14.6	14.4	14.5	4.8	8.4	6.6
Cutie [12]	44.8	**52.3**	**48.6**	74.8	79.5	77.2	**30.1**	**39.0**	**34.6**
SAM2 [64]	**45.2**	49.6	47.4	**87.5**	**90.8**	**89.1**	24.5	29.5	27.0

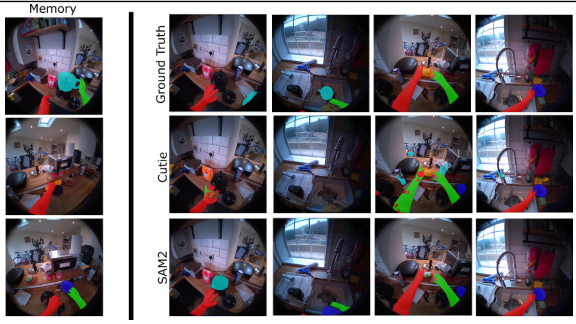

Figure 14. **Long-Term VOS**. jaccard index $\mathcal{J}$ & contour accuracy $\mathcal{F}$ show Cutie and SAM2 struggle with segmenting objects.

6. Onwards...

HD-EPIC is available from: http://dx.doi.org/10.5523/bris.3cqb5b81wk2dc2379fx1mrxh47 – *i.e.* the videos, audio, gaze, blender digital twin, camera pose estimates. Annotations are available at: http://hd-epic.github.io – *i.e.* object movements, object masks and 3D locations, long-object tracks, and object-action-fixture labels. We hope HD-EPIC will direct future research to a more holistic perception of egocentric videos.

Acknowledgements

Research at Bristol is supported by EPSRC Fellowship UMPIRE (EP/T004991/1), EPSRC Program Grant Visual AI (EP/T028572/1) and EPSRC Doctoral Training Program. O Emara, K Flanagan and F Abdelazim are supported by UKRI CD in Interactive AI (EP/S022937/1). The project is also supported by a unrestricted charitable donation from Meta (Aria Project Partnership) to the University of Bristol. Gemini Pro results are supported by a research credits grant from Google DeepMind.

Research at Leiden is supported by the Dutch Research Council (NWO) under a Veni grant (VI.Veni.222.160). Research at Singapore is supported by Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (No. MSS23C018).

We thank Rajan from Elancer and his team, for their huge assistance with temporal and audio annotation. We thank Srdjan Delic and his team for their assistance with mask annotations. We thank Damian Steer and the University of Bristol's RDSF team for hosting and maintaining the dataset. We also thank Owen Tyley for the 3D Digital Twin of the kitchen environments using Blender.

We thank David Fouhey and Evangelos Kazakos for early feedback on the project. We thank Pierre Moulon, Vijay Baiyya and Cheng Peng from the Aria team for technical assistance in using the MPS code and services.

We acknowledge the usage of EPSRC Tier-2 Jade clusters. The authors also acknowledge the use of Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

References

[1] Project aria machine perception services. https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps. 3, 2

[2] *Blender*. https://www.blender.org/. 5

[3] My fitness app. https://www.myfitnesspal.com/. 3, 1

[4] *VGG List Annotator (LISA)*, 2022. https://www.robots.ox.ac.uk/ vgg/software/lisa/. 5

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 9

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[7] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[8] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 2

[9] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. TIM: A Time Interval Machine for Audio-Visual Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7, 8, 15

[10] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. ReXTime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[12] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8, 15

[13] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. 2, 7, 14

[14] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 3

[15] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. TVBench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 2

[16] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide

Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 4

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)*, 2022. 2, 3, 4, 6, 5

[19] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 5, 7

[20] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the Carnegie Mellon University Multimodal activity (CMU-MMAC) database. 2009. 3

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 7

[22] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM)*, 2019. 5, 6

[23] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[24] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7, 15

[25] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 7

[26] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 15

[27] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-Supervised Audio Spectrogram Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 8

[28] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, 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, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 6

[29] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 6

[30] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional

neural networks and incremental parsing. 2020. 5, 10

[31] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, 2023. 4, 5

[32] Roland Johansson, Göran Westling, Anders Bäckström, and John Flanagan. Eye–Hand Coordination in Object Manipulation. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 2001. 5

[33] Amlan Kar, Seung Wook Kim, Marko Boben, Jun Gao, Tianxing Li, Huan Ling, Zian Wang, and Sanja Fidler. Toronto annotation suite. https://aidemos.cs.toronto.edu/toras, 2021. 7

[34] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-Fast Auditory Streams For Audio Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 8

[35] Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. ViLMA: A Zero-Shot Benchmark for Linguistic and Temporal Grounding in Video-Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5

[37] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 5

[38] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. NeMo: a toolkit for building AI applications using Neural Modules. *arXiv preprint arXiv:1909.09577*, 2019. 3

[39] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2023. 3

[40] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 1999. 5

[41] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[42] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[43] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[44] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[45] Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 3

[46] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 3

[47] Zhaoyang Lv, Nickolas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria Everyday Activities Dataset. *arXiv preprint arXiv:2402.13349*, 2024. 6, 3

[48] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[49] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[50] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria Digital Twin: A New Benchmark Dataset for Egocentric 3D Machine Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 3

[51] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7, 15

[52] Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Jikai Wang, Qifan Zhang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. CaptainCook4D: A Dataset for Understanding Errors in Procedural Activities. In

Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks, 2024. 3

[53] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 15

[54] Toby Perrett, Tengda Han, Dima Damen, and Andrew Zisserman. It's just another day: Unique video captioning by discriminitave prompting. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2024. 2

[55] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[56] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. In *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, 2025. 7

[57] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 15

[58] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[60] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023. 3

[61] Francesco Ragusa, Antonino Furnari, Sebastiano Battiato, Giovanni Signorello, and Giovanni Maria Farinella. EGO-CH: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. *Pattern Recognition Letters*, 2020. 3

[62] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3

[63] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. 2

[64] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 8, 7, 15

[65] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[66] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[67] Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons van der Sommen, et al. IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[68] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6, 3

[69] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[70] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[71] Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016. 3

[72] Kiran K. Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob J. Engel, Renzo De Nardi, and Richard A. Newcombe. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. *arXiv preprint arXiv:2308.13561*, 2023. 3, 1

[73] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4D Goal-Step: Toward Hierarchical Understanding of Procedural Activities. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[74] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. EFM3D: A Bench-

mark for Measuring Progress Towards 3D Egocentric Foundation Models. *arXiv preprint arXiv:2406.10224*, 2024. 6, 3

[75] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 7, 14

[76] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 7, 15

[77] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[78] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. SOK-Bench: A Situated Video Reasoning Benchmark with Aligned Open-World Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[79] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[80] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 3

[81] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[82] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? Visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[83] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 15

[84] Linjie Yang, Yuchen Fan, and Ning Xu. The 2nd large-scale video object segmentation challenge - video object segmentation track, 2019. 15

[85] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[86] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. MM-Ego: Towards Building Egocentric Multimodal LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[87] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 5

[88] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 2

[89] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 7, 14

[90] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7, 14

[91] Yunhan Zhao, Haoyu Ma, Shu Kong, and Charless Fowlkes. Instance tracking in 3D scenes from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[92] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle In A Video Haystack: A Scalable Synthetic Framework for Benchmarking Video MLLMs. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[93] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2