EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-centric View of Procedural Activities in Real World

Yifei Huang[†‡], Guo Chen[‡], Jilan Xu[‡], Mingfang Zhang[‡], Lijin Yang, Baoqi Pei
Hongjie Zhang, Lu Dong, Yali Wang[*♯], Limin Wang[*♭], Yu Qiao[*]
OpenGVLab, Shanghai AI Laboratory
♯Shenzhen Institutes of Advanced Technology, CAS ♭Nanjing University

Abstract

Being able to map the activities of others into one's own point of view is a fundamental human skill even from a very early age. Taking a step toward understanding this human ability, we introduce EgoExoLearn, a large-scale dataset that emulates the human demonstration following process, in which individuals record egocentric videos as they execute tasks guided by exocentric-view demonstration videos. Focusing on the potential applications in daily assistance and professional support, EgoExoLearn contains egocentric and demonstration video data spanning 120 hours captured in daily life scenarios and specialized laboratories. Along with the videos we record high-quality gaze data and provide detailed multimodal annotations, formulating a playground for modeling the human ability to bridge asynchronous procedural actions from different viewpoints. To this end, we present benchmarks such as cross-view association, cross-view action planning, and cross-view referenced skill assessment, along with detailed analysis. We expect EgoExoLearn can serve as an important resource for bridging the actions across views, thus paving the way for creating AI agents capable of seamlessly learning by observing humans in the real world. The dataset and benchmark codes are available at https://github.com/OpenGVLab/EgoExoLearn.

1. Introduction

Even as a child, humans can observe the actions of others and then map them to their own view [6, 41, 110, 115]. With this ability to asynchronously bridge activities from egocentric and exocentric views [105, 113], humans can watch others' demonstrations and replicate the procedures in a new environment. This ability is especially beneficial when actual physical trials carry the potential of high costs [31], *e.g.*, conducting dangerous chemical experiments.

In the wake of recent advancements in AI systems, one goal for the next generation of AI agents is to perform tasks in a more embodied setting [104]. However, different from humans, training these AI agents usually requires demonstration videos taken in a similar environment [84, 130] and from a congruent perspective with the AI agents, (*e.g.*, the egocentric point of view [50, 67, 118, 145]). While great effort has been made into the collection of egocentric data in different scenarios [22, 37, 116], it remains crucial for the AI agents to directly learn from demonstration videos taken in a different place and from a different viewpoint [42, 160]. Realizing this capability can unleash the full potential of public instructional video data [91] and is also useful in the human-robot cooperation scenario, especially in novel environments [65, 86, 139].

Current works towards this goal can be roughly divided into two directions. One way is to learn models in simulated environments [13, 73, 88, 99, 102], but it remains difficult for models in this setting to generalize in the real world [137]. The other direction is to learn from human activity in real-world scenarios. However, attempts to directly combine existing multiview datasets often yield datasets of inferior quality or scale [135, 154]. Meanwhile, the few existing datasets in this direction [109, 116, 119] only record ego- and exo-view videos in the same environment and in a time-synchronized manner. In reality, when following demonstrations it is often needed to bridge a series of procedural actions performed in a different place and at a different time. However, currently, no dataset is available for the exploration of how to bridge asynchronous procedural activities in realistic egocentric and exocentric viewpoints.

To address this lack of dataset issue, we introduce EgoExoLearn, a large-scale dataset containing demonstration videos and corresponding egocentric videos where the camera wearers follow the demonstrations and perform the same task in a different environment, as shown in Figure 1. Targeting two potential applications, *i.e.*, daily assistance and professional support, EgoExoLearn consists of 747 video sequences spanning a total of 120 hours of footage, ranging from daily food-making to specialized laboratory experi-

[*]Corresponding authors. [†]Project lead. [‡]Equal key contributions.

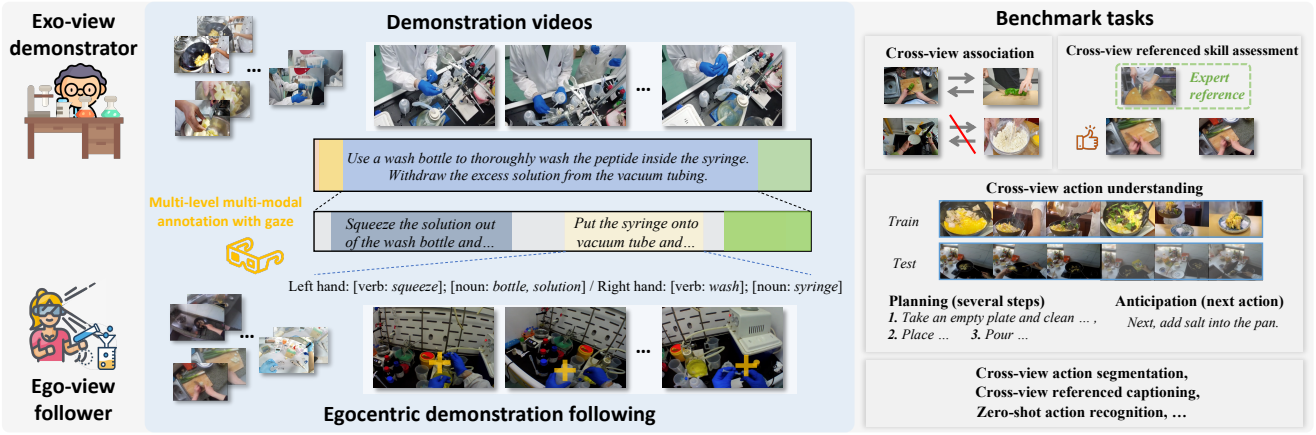


Figure 1. EgoExoLearn emulates the human asynchronous demonstration following process. It contains demonstration videos of multiple tasks, together with egocentric videos recorded by participants replicating the procedure after watching the demonstrations. The dataset comprises gaze signals and fine-grained multi-level multi-modal annotations, enabling the exploration of key features in this context such as cross-view association and cross-view action planning.

ments. Notably, the egocentric videos in EgoExoLearn contain eye gaze signals showing humans' visual attention while performing the task. This provides a valuable cue for better bridging the actions in ego- and exo-viewpoints.

We take one more step forward by analyzing human ability in bridging asynchronous ego- and exo-view actions and, accordingly, introduce new tasks and benchmarks that we believe can form building blocks for the development of next-stage embodied AI agents with similar abilities. When humans perform an action, he/she can associate and describe the undergoing action in the egocentric view with the corresponding action in the demonstration. With the knowledge from demonstration videos, humans can know the needed action steps and predict what the next steps should be. Besides, through the comparison with the demonstration, humans can also assess their level of skills.

Based on the above analysis, we design benchmarks of 1) cross-view association, 2) cross-view action understanding, 3) cross-view referenced skill assessment, and 4) cross-view referenced video captioning. Each benchmark is meticulously defined, annotated, and supported by baseline implementations. In addition, we pioneeringly explore the role of gaze in these tasks. We hope our dataset can provide resources for future work for bridging asynchronous procedural actions in ego- and exo-centric perspectives, thereby inspiring the design of AI agents adept at learning from real-world human demonstrations and mapping the procedural actions into robot-centric views.

2. Related Work

Ego-exo datasets. While there exist works that associate existing datasets to explore how activities can be bridged between them, these associated datasets are often limited in scale [17, 155, 159] or quality [135], meanwhile focusing only on single actions captured from the same view [90, 98, 142]. As for actions from different views, apart from multi-view fixed camera datasets [8, 19, 62, 70], there also exist datasets with both ego- exo-centric view videos [37, 38, 109, 116, 119]. These datasets are either recorded in the same environment [52, 109, 119] or record time-synced multi-view videos in the same environment with primary focuses on pose/activity understanding grounded in the 3D world [63, 116, 166]. Our dataset offers a more challenging and realistic scenario, where egocentric camera wearers learn to complete the tasks demonstrated by exocentric demonstration videos. This setting complements these datasets by focusing on high-level procedural actions.

The only dataset conceptually similar to ours is the recently proposed AE2 dataset [154], where the goal is to learn view-invariant representation from unpaired ego and exo videos. This dataset combines ego and exo videos from five public datasets [22, 24, 61, 63, 167] and a newly collected ego tennis forehand dataset. However, due to the difficulty in associating existing ego-exo datasets, the AE2 dataset is relatively small where the largest subset contains only 322 clips. Also, this dataset only focuses on clip-level actions, and thus cannot feature the real-world demonstration following setting, which usually requires multimodal, task-centric procedural knowledge. Instead, our EgoExoLearn is much larger in scale (100x more clips), while offering gaze and fine-grained multimodal annotations facilitating multi-faceted analysis of ego-exo action understanding.

Egocentric video datasets. In line with the recent development in wearable cameras [120], multiple egocentric video datasets [7, 21, 23, 37, 53, 87, 108, 122, 162] have been proposed. Different from previous egocentric datasets, the egocentric videos in EgoExoLearn feature a demonstration-following setting. We believe EgoExoLearn provides a playground for developing tools to bridge asynchronous procedural activities from ego- and exocentric viewpoints.

The setting of our EgoExoLearn complements existing datasets like Ego4D and can benefit from their rich knowledge and representations.

Egocentric gaze. Gaze can indicate visual attention and contains valuable information about human intent [43, 168], thus is used in a diverse range of areas such as human-computer interaction [49, 58, 163], and augmented reality [100, 112]. In computer vision, efforts have been made to leverage gaze in various tasks [44–46, 64, 74–76, 92, 111, 150]. However, with the previous absence of large-scale egocentric datasets that include gaze, this avenue of research is currently underexplored [51, 77, 104, 166, 172]. Our EgoExoLearn offers calibrated gaze positions for all egocentric videos. Thanks to our unique setting, our dataset enables the integration of gaze in egocentric video understanding and the exploration of the role of gaze in the cross-view context.

Egocentric and ego-exo video understanding. The unique recording perspective of egocentric videos presents a series of challenges including but not limited to action understanding [16, 33, 34, 47, 57, 103, 106, 123, 131, 132, 136], hand detection [14, 36, 117, 165], and video-language understanding [48, 55]. These form fundamental building block techniques of embodied AI [95], VR/AR [11, 54, 83, 134], and human-robot interaction [72, 89, 96, 148]. Since most egocentric datasets are smaller in scale compared with general datasets which contain mostly exocentric view videos [5, 153, 164], it is possible to leverage exocentric video data to improve model performance on egocentric videos [143]. There are typically three main directions: joint view-invariant learning [135, 151, 153], domain adaptation [147], and knowledge distillation [78]. In this work, we evaluate all these directions in our benchmarks.

3. Dataset

3.1. Data Collection

Scenarios and tasks. We consider procedural goal-oriented tasks ranging from daily food-making to specialized laboratory-based experiments. This selection is grounded in their exemplification of two prospective areas where future embodied AI agents would need the ability to bridge ego-exo activities: daily-life assistance and professional support. Specifically, EgoExoLearn incorporates 5 types of daily tasks (*e.g.*, cooking) and 3 types of specialized laboratory tasks (*e.g.*, solid-phase peptide synthesis). We record egocentric videos in 4 different kitchens and 3 different labs. Other details are provided in the supplementary.

Data collection procedure. Before the start of each collection session, participants are required to complete a questionnaire gathering basic demographic information and their self-evaluated expertise in executing the designated task. This questionnaire also highlights the ethical, privacy, and security considerations. Then in each session, participants will be asked to choose one or several exocentric view demonstration videos from a provided list and carefully learn the detailed procedures. Once they feel ready, they will wear Pupil Invisible Glasses [56], complete the gaze calibration, and begin to replicate the task performed in the demonstration videos. While not encouraged, participants are permitted to revisit the demonstration video during the recording.

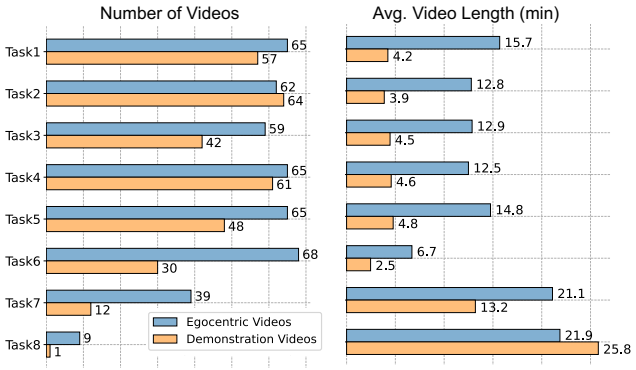


Figure 2. The number of videos per task (left) and the average duration of each video per task (right). Task1 to Task5 represent the 5 daily tasks and the remaining are three tasks in specialized laboratories. In each recording session of the egocentric video, one participant may learn from multiple demonstration videos and one demonstration video may be watched by several participants.

After each recording session, the participants are asked to re-do the gaze calibration to ensure gaze data fidelity. For the 5 daily tasks, the exocentric demonstration videos are manually curated from online video platforms such as YouTube. For the lab experiments, the exocentric demonstration videos are tutorials recorded by senior lab members.

Figure 2 shows the distribution of the 120 hours of data. Since most demonstration videos are meticulously edited to remove repeated steps, the average length of demonstration videos is lower than the egocentric videos which record the full procedure. As a result, the EgoExoLearn contains 432 egocentric videos totaling 96.5 hours and 315 demonstration videos spanning 23.5 hours. This difference in video length poses a unique challenge when bridging ego- and exo-view activities for future research endeavors.

3.2. Annotation

To facilitate our dataset in the development of algorithms that can effectively bridge the gap between ego and exo viewpoints, we provide detailed multi-modal human annotations. Our pipeline of annotation contains four stages detailed in the following paragraphs. Each of these stages is subject to a rigorous manual quality check involving no fewer than two individuals for verification and validation.

Coarse-level language annotation. In this step, we ask annotators to annotate the coarse actions in the videos. Like

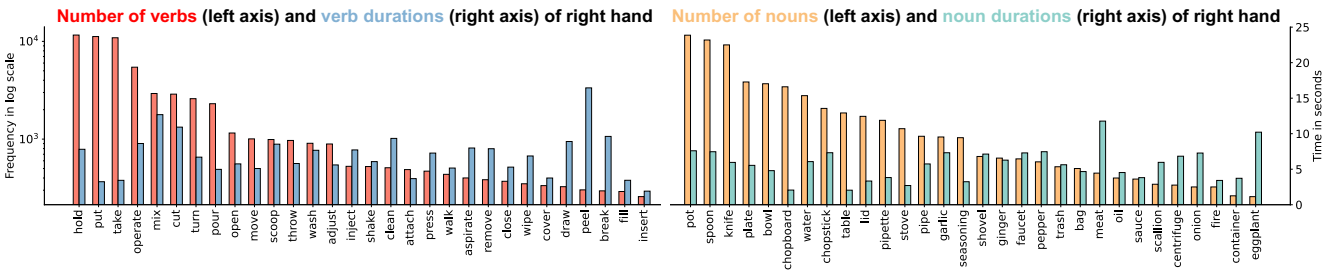


Figure 3. Occurrence and duration distribution of the annotated fine-level verbs and nouns associated with the right hand.

the previous works [116, 137], the coarse actions are defined as a middle-level step for accomplishing a task and can be divided into multiple fine actions. For instance, "Prepare the pork" in the task of twice-cooked pork, and "Suction filtration" in solid-phase peptide synthesis. Three types of annotation are given in this step: 1) the temporal interval, consisting of start and end timestamps; 2) the action label; and 3) a language description of the video within the annotated interval. We specifically request annotators to focus on elucidating "what is done," "how it is done," and "the purpose of this step" in their language descriptions. We define a total of 39 categories of coarse-level actions in this stage and acquire 41.2 coarse-level annotations per video with an average length of 21.5 seconds.

Fine-level language annotation. Based on the coarse-level annotations, in this step we request the annotators to provide annotations for the fine-level actions. The fine-level actions are the atomic actions like "take knife" or "pull syringe plunger". Unlike the first step, annotators are instructed to furnish language descriptions that specifically emphasize "which hand is used", "what object is used" and "why it is used". For the first two steps, we employ a two-round manual annotation checking to ensure the annotation quality.

Translation & parsing. To ensure linguistic precision, all the annotators give the language description using their native language [22]. For non-English annotations, we employ ChatGPT and Google Translation API to translate them into English. Subsequently, for the fine-level annotations, we employ specific rules and utilize tools such as NLTK [12] and Spacy [127] to extract the verbs and nouns associated with each segment. During this stage, we confine the selection of verbs and nouns to the predefined taxonomy offered by Ego4D [37], while also manually introducing supplementary verbs and nouns that are absent in the taxonomy. Since our manual annotation specifies the engagement of the left/right hand, we can extract multiple verbs and nouns for each segment, meanwhile attributing them to the respective involvement of the left or right hand. After manual checking, we obtain a total of 95 verb and 254 noun categories in the fine-level annotation. In Figure 3 we show the occurrence of the top 30 categories of verbs and nouns attributed to the right hand. More statistics can be found in the supplementary.

Skill level annotation. Since self-assessed skill level is not perfectly suitable for skill assessment, we identify several representative skills and assign human annotators to assess their skills. The annotation follows a pairwise ranking scheme, where annotators are presented with pairs of videos of the same action, and instructed to determine which video demonstrates a higher skill level. We prepare 40,191 video pairs and ensure that 4 different annotators annotate each pair. After filtering out pairs with less than 3 consistent opinions, we get a collection of 34,239 valid video pairs.

3.3. Statistics & Comparisons

To the best of our knowledge, there is no dataset that follows the same setting as ours for a direct comparison. Therefore, we enumerate various aspects of our dataset and conduct a comparative analysis with relevant datasets in Tables 1 and 2. EgoExoLearn distinctively enriches the domain with its "visual demonstration following" setting. Beyond this unique setting, it stands as the first egocentric dataset that includes temporal bounded language captions, annotated cross-view associations, and multi-label video segments.

4. Dataset Properties & Benchmarks

4.1. Dataset Properties

EgoExoLearn stands out from current egocentric and ego-exo datasets due to several unique properties.

Ego-Exo demonstration following setting. The most distinguished property of EgoExoLearn is the ego-exo demonstration following context. Egocentric video recorders are instructed to follow the steps in exocentric demonstration videos to perform the same task but in a different environment. This setting closely emulates the human observational learning process [6, 41] and can be instrumental in designing embodied AI agents that learn from alternative perspectives while executing tasks from their own viewpoint.

Fine-grained vision-language annotations with gaze. To facilitate a deeper analysis, we equip our dataset with rich, multimodal, fine-grained annotations. EgoExoLearn is the first egocentric dataset featuring high-quality captions, evidenced by the number of words per segment in Tab. 2.

Dataset	Settings	Unique Hours	Ego +Exo?	Instruction following?	Visual instruction?	Gaze	Coarse Action	Fine Action	Dense Caption	Association	Skill
Meccano [108]	Industry	7	✗	✓	✗	✗	✗	✓	✗	✗	✗
EGTEA [76]	Cooking	28	✗	✓	✗	✓	✗	✓	✗	✗	✗
EK-100 [22]	Cooking	100	✗	✗	✗	✗	✗	✓	✗	✗	✗
HoloAssist [137]	Assistive	166	✗	✓	✗	✓	✓	✓	✗	✗	✓
Ego4D [37]	Multiple	3670	✗	✗	✗	◇	◇	◇	◇	✗	✗
H2O [63]	Desk	1	✓	✗	✗	✗	✗	✓	✗	✗	✗
LEMMA [52]	Daily	10	✓	✗	✗	✗	✗	✓	✗	✗	✗
HOMEAGE [109]	Daily	25	✓	✗	✗	✗	✓	✓	✗	✗	✗
CharadesEgo [119]	Daily	34	✓	✗	✗	✗	✗	✓	✗	✗	✗
Assembly101 [63]	Desk	42	✓	✓	✗	✗	✓	✓	✗	✗	✗
EgoExoLearn (ours)	Daily & Lab	120	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. **Comparison to related datasets on settings (left) and annotations (right).** "Unique hours" refers to the cumulative duration of distinct video recordings, counting only one camera's footage for simultaneous recordings of the same activity. ◇: Partially included.

Dataset	# videos	Avg. min	# segs	Avg. sec	verb classes	noun classes	# verb / seg	#noun / seg	#words / seg
CharadesEgo [119]	7860	0.5	69k	1.8	32	37	1	1	0
HOMEAGE [109]	5700	0.9	26k	-	29	86	-	-	0
EK-100 [22]	700	8.5	90k	3.1	97	300	1	1.2	3.0
Assembly101 [116]	4321	7.1	83k	1.7	24	90	1	1	0
Ego4D [37]	991	26.4	77k	8.0	115	478	1	1	7.4
Ours-ego	432	13.4	64k	4.6	95	254	1.8	2.5	16.9
Ours-exo	315	4.5	14k	4.7	82	251	2.3	3.0	19.4

Table 2. **Contemporary egocentric datasets.** We show only the fine-level actions for a fair comparison. For Ego4D [37], we select the closest subtask of "forecasting" following [116].

Different from Ego4D where captions are associated with only single timestamps, our captions come with manually annotated start and end timestamps. For better visual perception across ego-exo views, we give verb and noun labels associated with the specific hand. One aspect from which we can analyze the human's ability to bridge ego-exo activities is through gaze. EgoExoLearn is also augmented with calibrated eye-gaze signals. These annotations enable the understanding of human ability to bridge ego-exo activities from diverse perspectives, which we posit will benefit the next-generation embodied AI agents [25, 71].

4.2. Benchmarks

To evaluate the ability of bridging asynchronous ego-exo procedural activities, we introduce 4 new benchmarks: 1) cross-view association, 2) cross-view action understanding, 3) cross-view referenced skill assessment, and 4) cross-view referenced captioning. The cross-view action understanding benchmark is further subdivided into three subtasks: cross-view action anticipation, cross-view action planning, and cross-view action segmentation. Additionally, we explore the role of gaze in assisting these tasks. We also benchmark models on zero-shot and supervised fine-grained action recognition tasks for reference, following [22, 137]. Note that we carefully split our dataset to eliminate annotation leak across benchmarks. Due to the space limit, we only provide partial content of definition, annotation, results, and analysis, leaving more complete details in the supplementary.

4.2.1 Cross-view association

Motivation. One straightforward indicator of the ego-exo activity bridging is the ability to associate the same semantics across ego- and exo-views. This benchmark focuses on equipping models with this cross-view association ability. An application of this ability is assistants in AR that can show expert demonstration videos when the human is confused [169]. Another potential application is the embodied AI agent that can explain its decision [129, 138, 158].

Problem settings. We formulate this association benchmark as a cross-view multiple-choice association problem. Specifically, we consider two different cross-view association settings: ego2exo and exo2ego. In the case of ego2exo, given an egocentric video, the model is asked to predict the corresponding exocentric video performing the same action from a candidate choice set of exocentric samples, and vice versa for the exo2ego setting. For both ego2exo and exo2ego settings, we use 20 candidate samples for each query. The evaluation metric is the averaged Top-1 accuracy.

Annotations. We meticulously construct the ground-truth ego-exo pairs via a semantic-aware matching process. It is composed of five stages with details in the supplementary material: (1) Scenario Matching. (2) Noun and Verb Matching. (3) Sentence Matching with LLM. (4) Negative Sampling. (5) Two-round Manual Verification. Notably, we do not provide such pairs for the training set and leave the modeling of cross-view association on unpaired samples to be further explored for the community.

Baseline model. We adopt three types of baseline models, *i.e.*, ego-only, exo-only, and ego-exo, which refer to the data we used during training. Under all the settings, we leverage the paired video and caption and jointly train a video encoder and a text encoder using the contrastive loss [107]. We use TimeSformer-B [10] as the video encoder and clip-text [107] as the text encoder. We initialize both encoders

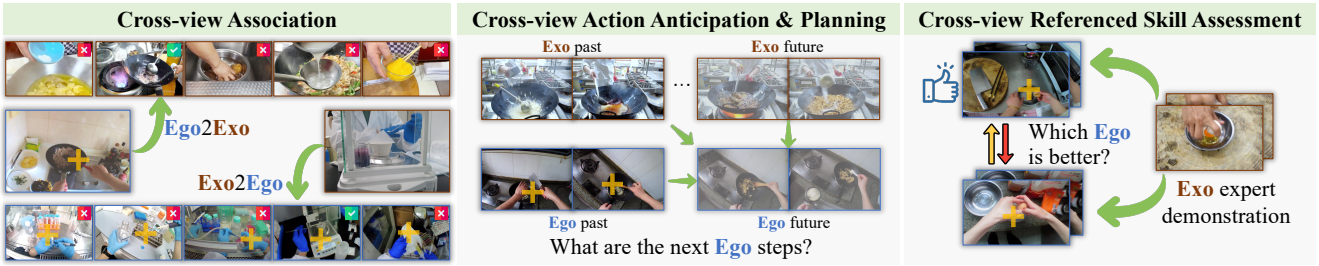


Figure 4. Concept of the 3 benchmarks of cross-view association (Sec. 4.2.1), cross-view action anticipation & planning (Sec. 4.2.2) and cross-view reference skill assessment (Sec.4.2.3) in this section. Other benchmarks can be found in the supplementary material.

using the EgoVLP [81] pre-trained weights. During testing, we obtain the ego/exocentric video representations using the video encoder. The prediction is defined as the one with the highest normalized cross-view video feature similarity among all candidates. On top of these models, we evaluate the effectiveness of gaze in associating egocentric video and exocentric video. This is achieved by replacing the original egocentric video with spatially cropped videos using the gaze positions as the cropping center.

Related work. Prior works on pre-trained vision-language models for multiple choice association/questioning [68, 81, 133, 141, 149, 170] are generally pre-trained on either exocentric [5, 91] or egocentric [37] datasets, only revealing weak cross-view bridging ability. Another line of works explores cross-view learning [3, 40, 146, 152] by either transferring the knowledge from one view to the other [78] or training view-invariant video understanding models [135, 154]. Different from previous works, our cross-view association benchmark is in a more realistic but challenging setting, by evaluating the model's ability to associate asynchronous activities across ego- and exo-views.

Experiment results. We first evaluate several vision-language models via zero-shot transfer. These models are pre-trained either on egocentric videos, *i.e.* EgoVLP [81] and LaViLa [170], or exocentric videos, *i.e.,* InternVideo [140]. As shown in Tab. 3, without using gaze, EgoVLP generally outperforms the others on both validation and test sets. By introducing gaze information, InternVideo receives a decent improvement, especially on Exo2Ego.

As for fine-tuned models, Tab. 3, reveals that models trained solely on single-view data struggle with cross-view association. Training ego-only models using gaze-cropped egocentric videos results in substantial improvements, significantly outperforming those trained on center-cropped videos. This again highlights the importance of gaze in enhancing cross-view association. Based on our observation, the regions around gaze help the cluttered ego videos become visually similar to exo videos where the primary object is salient. Last, we show the baseline result when co-trained on both egocentric and exocentric videos. The model in this setting shows the strongest cross-view association ability over single-view models. Findings from this benchmark underline the limitation of current models in associating activities across ego and exo views, and point towards the potential benefits of integrating gaze into the association.

Method	Gaze	Val		Test	
		Ego2Exo	Exo2Ego	Ego2Exo	Exo2Ego
Zero-shot					
Random	✗	12.7	15.0	14.1	13.4
EgoVLP [81]	✗	28.8	27.2	32.1	28.9
LaViLa [170]	✗	22.6	24.9	28.7	25.7
InternVideo [140]	✗	27.0	21.2	30.6	21.7
EgoVLP [81]	✓	28.8	29.7	31.5	28.9
LaViLa [170]	✓	21.9	21.4	30.3	25.9
InternVideo [140]	✓	**30.9**	**32.3**	**33.3**	**32.2**
Fine-tuned					
Exo-only	✗	42.9	41.7	45.4	46.9
Ego-only	✗	33.6	37.1	40.3	35.8
Ego-only	✓	34.6	38.7	45.6	41.8
Ego-only	Center	25.4	22.8	24.7	24.2
EgoExo	✗	42.9	45.4	49.0	45.3
EgoExo	✓	**47.9**	**48.8**	**55.3**	**51.1**

Table 3. Association accuracy in the cross-view association benchmark. In the *fine-tuned* setting, we adopt three kinds of data sources for training, *i.e.*, ego-only, exo-only, and hybrid ego-exo data. By leveraging gaze information during training, the model outperforms the baseline (w/o gaze) and the center-crop counterpart.

4.2.2 Cross-view action anticipation & planning

Motivation. The procedural actions are not guaranteed to be identical between the two views due to practical constraints. Thus, we design benchmarks for cross-view action anticipation and planning to enable a thorough understanding and transfer of necessary steps (or actions) for task completion, bridging the gap between views and considering real-world conditions. A practical application of this benchmark can be seen in human-robot collaboration scenarios. For instance, an embodied AI agent, after observing a human perform the first part of a task, could effectively take over and complete the remaining half of the task based on the particular environmental situation.

Problem settings. For both the cross-view anticipation and

Method	Gaze	Anticipation↑				Planning↓	
		Ego-V	Ego-N	Exo-V	Exo-N	Ego	Exo
Exo-only	✗	29.9	23.6	**40.9**	**40.5**	84.7	**76.1**
Ego-only	✗	33.4	37.8	28.9	17.6	83.4	84.5
Ego-only	✓	**40.5**	**52.8**	37.6	37.6	**80.0**	82.6
Ego-only	Center	33.2	38.6	34.1	32.7	82.6	84.7
Unsupervised Domain Adaption							
Ego2Exo	✗	33.6	38.1	35.4	28.7	83.0	84.1
Exo2Ego	✗	30.4	23.6	**39.2**	39.8	83.9	79.0
Ego2Exo	✓	**40.8**	**54.2**	38.7	37.1	82.8	84.3
Exo2Ego	✓	33.5	31.3	39.1	**40.1**	**82.4**	**78.8**
Knowledge Distillation							
Ego2Exo	✗	29.6	24.9	**41.6**	**45.2**	84.3	75.5
Exo2Ego	✗	34.0	38.4	28.6	18.6	83.1	84.3
Ego2Exo	✓	29.9	25.0	41.2	45.1	84.8	**75.1**
Exo2Ego	✓	**41.0**	**56.1**	37.7	39.1	**79.5**	82.6
Co-training							
Ego & Exo	✗	33.5	37.4	39.6	44.3	83.2	76.0
Ego & Exo	✓	**43.8**	**53.3**	**40.3**	**44.4**	**79.0**	**75.6**

Table 4. Results of cross-view action anticipation and planning benchmarks. For anticipation, the class-mean Top-5 recall is used as the evaluation metric (higher is better). For planning, the Edit distance is used as the evaluation metric (lower is better).

cross-view planning, the goal is to anticipate the future activities in one view, given labeled training data only in another view. For the cross-view anticipation task, we focus on predicting the verb and noun categories of the next fine-level action $\tau = 1$ second into the future. Given the multilabel nature of our verb and noun annotations in each fine-level segment, we perform multiclass anticipation. Performance is evaluated by class-mean Top-5 recall as per [22]. For the cross-view planning task, we aim to generate the next $K = 8$ steps of coarse-level actions. We adopt ED@K as the evaluation metric following the setting of Ego4D LTA [37].

Annotations. For action anticipation, we use the fine-level verb and noun annotations, and then take their intersection between ego and exo videos to constrain them in the same closed set. To evaluate the model more effectively, we further control the long-tail degree of the data by and filter out the tail categories that occur less than 1/100 of the highest occurrence category. For the action planning task, we directly adopt the coarse-level action annotations and regard the start timestamp of each segment as one action step. More details can be found in the supplementary material.

Baseline model. We explore three distinct directions to implement cross-view baseline models. The first direction is based on unsupervised domain adaptation (UDA), treating one view as the source domain and the other as the target domain. This method operates within an unsupervised training framework, using labels from the source domain and video data from both domains [20, 59, 94, 114, 121, 142, 157]. We adopt CLIP [107] + TA3N [17] as the baseline model. The second direction entails knowledge distillation (KD) [35, 39, 97], allowing the model trained on one view to learn knowledge of the other view, under the assumption that a teacher model of the other view is available. We equip CLIP with a distillation approach based on [78] to transfer knowledge from the teacher model to a newly created student model. The third and most straightforward direction is co-training (CT) using the data from both views to encourage the model to discover correlations between them directly. For using gaze, we also crop the video based on the gaze positions.

We comprehensively consider four evaluation settings. The "Ego-only" and "Exo-only" settings do not involve cross-view understanding, thus we use zero-shot evaluation serving as the references. The Ego2Exo and Exo2Ego settings are the cross-view settings. For UDA, "Ego2Exo" is defined as utilizing the egocentric view as the source domain and the exocentric view as the target domain. In the context of KD, "Ego2Exo" indicates we initially train a teacher model on egocentric data, followed by the training and distillation of the student model on exocentric data. For CT, we merge both egocentric and exocentric datasets through direct concatenation. We report results on the test set and put the validation set results in the supplementary.

Related work. Prior works [32, 62, 116, 128] on multi-view action understanding mainly focus on synchronized multi-view videos. Some work studied transferring knowledge [78] from one view to the other or training the view-invariant [135, 154, 171] video models. Our cross-view benchmarks seek to evaluate the ability of models to bridge asynchronous actions across views, which is more challenging yet realistic.

Experiment results. Table 4 presents the results of action anticipation and planning on the test set. The first block of results shows a significant performance gap when models trained exclusively on one view are tested on the other view. This underscores the inherent differences in activities captured in the two views. Remarkably, even without relying on any specific cross-view method, the inclusion of gaze information markedly diminishes the disparity between egocentric and exocentric data. Leveraging techniques such as UDA or KD we can see improved performance compared with direct zero-shot inference in the first block. Since CT can utilize both egocentric and exocentric data, it achieves the best performance in all setups. Moreover, from the comparison in all settings using or not using gaze, it is clear that gaze serves as a surprisingly effective signal to mitigate the gap between activities in the two views, although naively designed. These results take the first step in the potential directions for better bridging the cross-view activities, setting a foundation for future advancements in the field of cross-view action understanding.

4.2.3 Cross-view referenced skill assessment

Motivation and problem setting. We propose a novel task of cross-view referenced skill assessment leveraging the

unique setting and annotations of our dataset. This task goes beyond the traditional pairwise ranking often used in skill assessment [26, 66] by incorporating an expert demonstration video as a reference point. This demonstration provides a model of the ideal execution of an action, offering a standard against which to compare skill levels. In this task, the input is a pair of egocentric video clips C_{ego1}, C_{ego2} of the same action and an exo-view demonstration video clip C_{exo} as the reference, and the output is a choice $c \in \{ego1, ego2\}$ of the egocentric clip that demonstrates a higher skill level. Moreover, we can also assess the skill level by the relation between action and gaze. This benchmark evaluates a model's ability to bridge asynchronous ego-exo dynamics. A practical application is an AR system that helps humans in skill acquisition by providing targeted feedback based on skill level assessment and expert demonstration.

Annotations. For the cross-view referenced skill assessment benchmark, we concentrate on four types of representative actions, as detailed in Tab. 5. We use the skill level annotations in Sec. 3.2 in this benchmark. For each pair of videos, the annotation is provided by four distinct annotators to minimize subjective bias. We ensure credibility by checking the transitivity of annotations and removing the pairs with less than 3 agreements among 4 annotators. We then append an exo-view demonstration video clip of the same action, forming video triplets as the model input.

Baseline model. Our baseline model is built upon a pairwise ranking skill assessment model RAAN [27]. Without loss of generality, assuming $ego1$ is the video showing a higher skill level, we apply the following approaches to leverage the reference exo-view demonstration video: 1) Triplet loss (TL). The feature distance between C_{exo} and C_{ego1} should be closer to the distance between C_{exo} and C_{ego2}. 2) Relation network (RN). Inspired by [124], we employ a relation network that concatenates the features of the ego and exo clips. This network is designed to discern which of the two egocentric video clips bears a closer relation to the demonstration video in terms of skill level. The way to gaze is consistent with the other benchmarks.

Related work. Several previous works on skill assessment aim to directly regress a score based on professional ratings [1, 80, 82, 101, 125, 161]. We adopt a more general approach of pairwise ranking since no absolute score is available in most real-world skills [9, 79]. Previous works in this direction only use a pair of videos that are in either egocentric view [26] or in exo view [27, 79]. Differently from their works, we explore the cross-view activity bridging ability by examining how demonstration videos from exo view can benefit skill assessment in ego views.

Experiment results. We first evaluate the performance of previous works under the conventional pairwise setting. In the upper block of Table 5, both methods [26, 27] receive a clear performance boost when gaze is used. This aligns with behavioral science findings that experts and novices have different gaze patterns in the same task [85, 144]. With the exocentric demonstration as reference, both the relation network method (RN) and the triplet loss method (TL) can leverage the reference video and bring performance improvement. Further adding gaze we can get the best performance. However, the marginal improvement observed with the inclusion of the exocentric reference suggests that current models may still struggle to fully bridge asynchronous activities across egocentric and exocentric views. There remains ample room for new improvement and innovation.

Method	Gaze	Egg Cracking	Peeling	Stir-fry	Cutting
Ego pairs only					
Who's better [26]	✗	74.79	75.98	77.54	76.85
RAAN [27]	✗	78.45	78.52	82.53	79.09
Who's better [26]	✓	77.91	76.70	79.13	77.02
RAAN [27]	✓	82.08	79.37	83.92	79.36
Ego pairs + Exo					
RAAN [27] + RN	✗	77.92	77.09	81.54	78.26
RAAN [27] + TL	✗	78.81	79.46	82.50	78.73
RAAN [27] + RN	✓	82.14	79.32	83.51	**79.44**
RAAN [27] + TL	✓	**82.16**	**79.59**	**84.05**	79.29

Table 5. Ranking accuracy of cross-view referenced skill assessment. In the upper part of the table, only ego video pairs are used, while in the lower part, exo demonstrations are incorporated by "RN": relation network and "TL": triplet loss.

5. Conclusion

The ability to bridge asynchronous procedural activities in ego- and exo-views is imperative for next-generation embodied AI in executing sophisticated tasks in the real world. As a fundamental step, our `EgoExoLearn` encompasses a rich collection of egocentric videos, each captured when replicating procedures of exocentric demonstration videos, but performed in different environments and at different times. This realistic setup, combined with our multimodal annotations, allows us to construct 4 novel benchmarks, serving as a versatile platform for investigating how cross-view asynchronous activities can be bridged. `EgoExoLearn` also enables new research directions *e.g.,* how to better leverage gaze and hand-associated annotations. Results from the benchmarks show weaknesses of current models in bridging ego- and exo-view asynchronous activities, leaving significant room for future work to improve upon.

Acknowledgement. This work is supported by the National Key R&D Program of China (No.2022ZD0160102) and the Industry Collaboration Projects Grant, Shanghai Committee of Science and Technology, China (No.22YF1461500).

Key contribution statement: Guo made key contributions to annotation processing and 3 action benchmarks. Jilan made key contributions to the association and caption benchmarks. Mingfang contributed primarily to the skill benchmark.

EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-centric View of Procedural Activities in Real World

Supplementary Material

This supplementary material shows details about our benchmark including formal definition, implementation, and additional experiment results. Also, we show additional details about the collection and annotation of the dataset.

S6. Additional Benchmark Details

S6.1. Cross-view association

S6.1.1 Detailed task definition

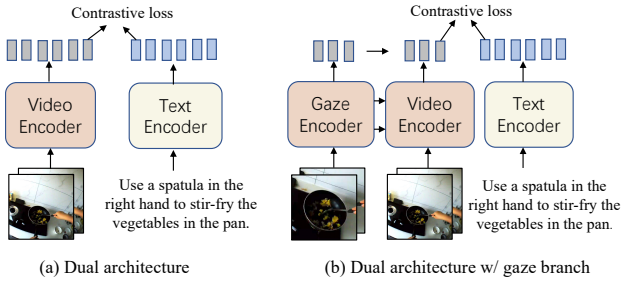
The training set consists of separate egocentric videos V^{ego} with associated narration T^{ego} and exocentric videos and narrations (V^{exo},T^{exo}). For each egocentric video, a sequence g with corresponding gaze is provided. Note that, we do not provide explicit pair information in the training set.

In the validation/test set, we introduce two evaluation settings, *i.e.,* Ego2Exo and Exo2Ego. We describe the formulation of Ego2Exo as follows. Each sample consists of an egocentric query video V^{ego} and K exocentric candidate videos $\{V_1^{\text{exo}}, ..., V_K^{\text{exo}}\}$, where only one candidate exocentric video corresponds to the query egocentric video, *i.e.,* the same action is being performed. In the Exo2Ego setting, the query is exocentric videos while egocentric videos form the candidate set. For both Ego2Exo and Exo2Ego settings, we consider $K = 20$ candidates.

S6.1.2 Implementation details

Training setting. As explicit pairing is not available in the training set, we propose a simple baseline approach to align egocentric videos and exocentric videos in the semantic space. In specific, we train a dual-encoder architecture consisting of a video encoder $f_v(\cdot)$ and a text encoder $f_t(\cdot)$ on both ego- and exo-videos and narrations using the contrastive loss, named as *co-training* in our experiments. Following [81, 170], we adopt a TimeSformer-B [10] as the video encoder and a clip [107] text encoder. We randomly sample 4 frames as input. The model is initialized with weights pre-trained on Ego4d video and text pairs [37, 81]. We train the dual encoder model for 5 epochs with a fixed learning rate 1e-5 and a batch size of 32. At the inference stage, the text encoder is discarded and only the video encoder is used. For each query, we compute its video representation with K features of the candidate videos and select the one with highest cosine similarity as the model prediction.

Network architecture. To leverage the gaze information in associating egocentric and exocentric videos, we further propose a multi-view branch for the video encoder [156]. One branch encodes the original video while the other branch encodes the gaze cropped video, as illustrated in Fig. S5. The feature of the original video cross-attends to the gazed video feature every at the 5$^{\text{th}}$, 8$^{\text{th}}$, and 11$^{\text{th}}$ transformer block, enabling multi-scale feature fusion for improved visual representation. For exocentric videos, we simply input the original video to the gaze branch.



(a) Dual architecture (b) Dual architecture w/ gaze branch

Figure S5. Cross-view association network with naive dual architecture (a) and improved architecture with additional gaze branch (b).

S6.1.3 Annotation details

The process of pair construction consists of five stages: (1) Scenario Matching. We gather all the egocentric and exocentric videos under the same scenario (e.g. cooking the same dish or conducting the same experiment) into each group. (2) Noun and Verb Matching. Based on the noun and verb vocabularies, for each group, we pair an egocentric caption with another if they contain exactly the same nouns and verbs. (3) Sentence Matching with LLM. We ask the LLM (e.g. ChatGPT) to determine whether each ego-exo caption pair obtained in stage 2 describes the same activity at sentence-level, reducing the linguistic ambiguity caused by word matching. (4) Negative Sampling. We randomly choose video clips from the same video as negative samples in the candidate set. (5) Two-round Manual Verification. We manually check the semantic meaning of each ego-exo pair and corresponding ego-exo video to make sure the exact match. This verification is performed in two rounds by two different individuals. In total, the size of the validation/test set is 868/2200. As stated in the main manuscript, we do not provide such pairs for the training set and leave the modeling of cross-view association on unpaired samples to be further explored for the community.

S6.2. Cross-view action anticipation & planning

S6.2.1 Detailed task definition

Task definitions of cross-view action anticipation and planning have followed the previous benchmarks of [22] and [37]. Our cross-view benchmark extends on the original task setting and focuses on mutual assistance between egocentric and exocentric video data.

Action anticipation. The action anticipation task focuses on forecasting the verb and noun categories of the subsequent fine-level action at $\tau = 1$ second into the future. Considering a fine-level action segment $a = (s, e, c)$, where s, e, and c represent the start time, end time, and category of a respectively, the model is restricted to observing video data only up to time $s - \tau$. The model's objective

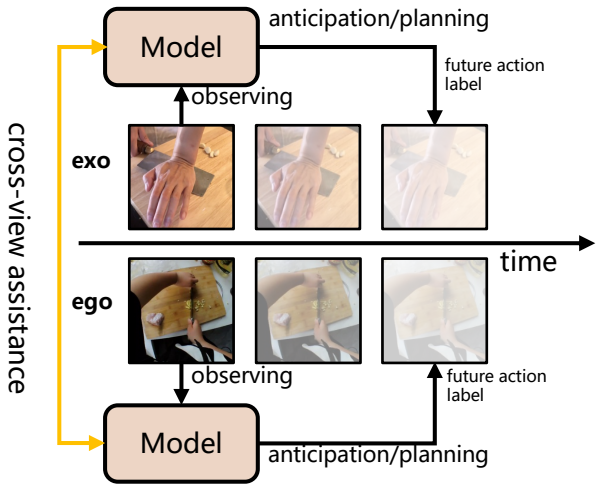


Figure S6. Overall framework of cross-view action anticipation and planning. The model observes the past video and tries to anticipate the next fine-level action (action anticipation) or the next K steps of the coarse-level actions (action planning). The model gets assistance from the knowledge in the other view.

is to predict the forthcoming action, encompassing relevant verbs and nouns. The performance of the model in this benchmark is evaluated using class-mean Top-5 recall, as outlined in [22].

Action planning. The objective of the action planning task is to generate the next K steps of coarse-level actions. Considering N_a fine-level action segments $A = \{a_i = (s_i, c_i)\}_{i=1}^{N_a}$, where s_i (ensuring $s_i < s_{i+1}$) and c_i represent the start time and category of a_i respectively, the model is limited to observing video data up to time s_i and is tasked with forecasting the K actions $s_i, ..., s_{i+K-1}$ into the future. For evaluation purposes, we adopt ED@K as the metric, following the approach outlined in Ego4D LTA [37]. In our specific configuration, we set K to 8 and sample 5 predicted sequences for evaluation.

Cross-view benchmark. In our cross-view benchmark, we begin by assessing zero-shot cross-view action understanding. Following this, we employ various methods to leverage information in one view to assist the understanding in the other view. Thus, this benchmark is focused on designing approaches that utilize both ego and exo-view data to enhance the cross-view performance. Figure S6 shows the overall framework of our cross-view benchmark for action anticipation and planning. Figure S7 further illustrates our various cross-view settings.

S6.2.2 Implementation details

Network architecture. To adapt our cross-view training settings, we rely on the TA3N [17] code base, acknowledged for its clarity and comprehensibility, and widely adopted in recent research. We employ CLIP [107] as the feature extractor for generating frame-level video features. Both action anticipation and planning tasks entail leveraging historical information to forecast future actions. Thus, we input a 2-second context into the model. Within the specified temporal range, we uniformly sample 5 frames as the input. Utilizing the 3D feature map extracted by TA3N [17], we perform average pooling to condense the feature map into a vector $v \in \mathbb{R}^d$. We employ a projector $\mathbf{W}_{anti}$ to predict C_{anti} classes for the action anticipation task, where C_{anti} is the number of verb or noun categories. For the action planning task, we use a projector $\mathbf{W}_{plan}$ to predict $C_{plan} \times K$ classes, where C_{plan} is the number of coarse-level categories and K (set to 8) is defined in Sec S6.2.1.

Training. We first introduce the training settings of both tasks. Given the anticipation logits y_{anti} produced by the model and the corresponding ground truth $\hat{y}_{anti}$, we employ the standard cross-entropy loss for supervision:

$$\mathcal{L}_{anti} = \mathcal{L}_{CE}(y_{anti}, \hat{y}_{anti}). \tag{1}$$

For an action sequence $y^1_{plan}, \ldots, y^K_{plan}$ predicted by the action planning model, the loss function is defined as:

$$\mathcal{L}_{plan} = \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_{CE}(y^i_{plan}, \hat{y}^i_{plan}). \tag{2}$$

The model is trained using the SGD optimizer with a learning rate set to 1e-2 and the training process spans 40 epochs.

Zero-shot cross-view setting. In the zero-shot cross-view setting, the model is initially trained on data in one view and directly tested on data in the other view. This setting is crucial for understanding how well a model trained on data from one perspective can adapt to and accurately interpret data from another perspective, without any additional training specific to that new viewpoint. Figure S7(a) illustrates the procedure of the "exo2ego" cross-view setting, where the model is first trained on exocentric data and then tested on egocentric data. The "ego2exo" setting works vice versa.

Unsupervised domain adaptation setting. In the unsupervised domain adaptation setting, the training process involves using data and labels from the source view, plus the video data from the target view. The annotations from the target view are not used. Figure S7(b) illustrates the "exo2ego" cross-view setting, where exocentric data serves as the source domain, and egocentric data serves as the target domain. In addition to task supervision, the overall loss function also contains a domain adaption loss derived from TA3N [17] for unsupervised domain adaptation settings:

$$\begin{aligned}\mathcal{L}_{DA} = &\frac{1}{N_S}\sum_{i=1}^{N_S}\mathcal{L}^i_y + \frac{1}{N_{S\cup T}}\sum_{i=1}^{N_{S\cup T}}\gamma\mathcal{L}^i_{ae} \\ &- \frac{1}{N_{S\cup T}}\sum_{i=1}^{N_{S\cup T}}(\gamma^s\mathcal{L}^i_{sd} + \gamma^r\mathcal{L}^i_{rd} + \gamma^t\mathcal{L}^i_{td}).\end{aligned} \tag{3}$$

Therefore, the overall loss function for both tasks under this setting is

$$\mathcal{L} = \mathcal{L}_{anti/plan} + \mathcal{L}_{DA}. \tag{4}$$

Knowledge distillation setting. In the knowledge distillation setting, the training process comprises two stages: (1) training the teacher model on the data in one view, and (2) training the student model on the data in the other view, meanwhile distilling knowledge from the teacher model. Figure S7(c) depicts the "exo2ego" cross-view setting, where the teacher model is trained on exocentric data, and the student model is trained on egocentric data. In addition to task supervision, the overall loss function for training

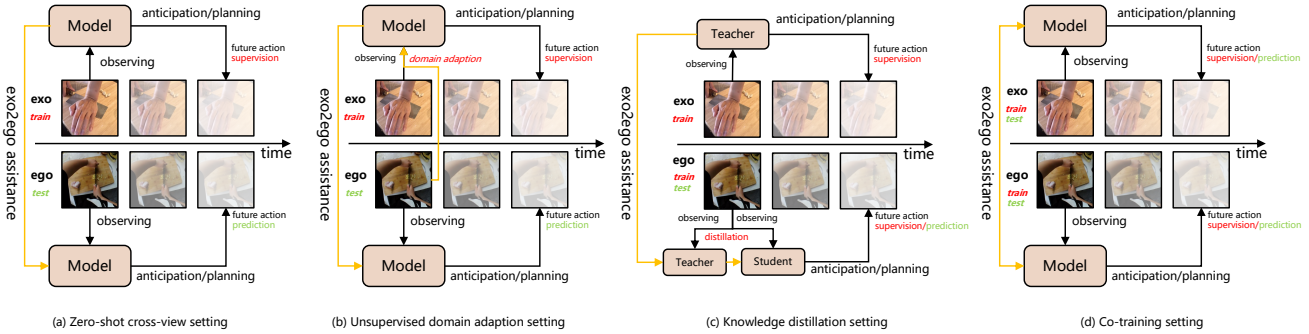


Figure S7. Four settings for cross-view action anticipation and cross-view action planning. (a) The zero-shot setting directly evaluates the model trained on one view on the test data of the other view. (b) The unsupervised domain adaptation (UDA) setting involves leveraging data from another view, but without using the labels associated with this data. (c) In the knowledge distillation setting, for a model in one view, a teacher model trained on the other view is used to provide assistance. (d) The co-training setting directly uses the data and labels of both views. (a) to (d) represent 4 increasing degrees of cross-view information usage.

the student model also contains a knowledge distillation loss for knowledge distillation settings. Specifically, we use L2 loss to minimize the feature y^S_{feat} and y^T_{feat} output by the student and teacher:

$$\mathcal{L}_{KD} = \mathcal{L}_{L2}(y^S_{feat}, y^T_{feat}). \tag{5}$$

Therefore, the overall loss function of the teacher and student for both tasks under this setting is

$$\mathcal{L}_{teacher} = \mathcal{L}_{anti/plan}, \tag{6}$$
$$\mathcal{L}_{student} = \mathcal{L}_{anti/plan} + \mathcal{L}_{KD}. \tag{7}$$

Co-training setting. In the co-training setting, exocentric and egocentric data are both used to train the model. The model is then evaluated on the test set of the egocentric and exocentric data. Figure S7(d) depicts the "exo & ego" co-training setting.

S6.2.3 Annotation details

Cross-view action anticipation. The annotation process for cross-view action anticipation involves three stages: (1) extracting verbs and nouns for each fine-level action clip, (2) aligning the closed categories of training, validation, and testing set across egocentric and exocentric videos, (3) restricting the closed set to the intersection of categories present in all egocentric and exocentric videos, and (4) managing the long-tail distribution of the data by filtering out categories that occur less than 1/100 of the highest occurrence category. We delete all video clips without any label. As a result, this task contains 19/31 verb/noun categories. The size of the egocentric train/validation/test set is 34.5k/7.7k/17.3k, and the size of the exocentric train/validation/test set is 6.1k/2.1k/4.8k.

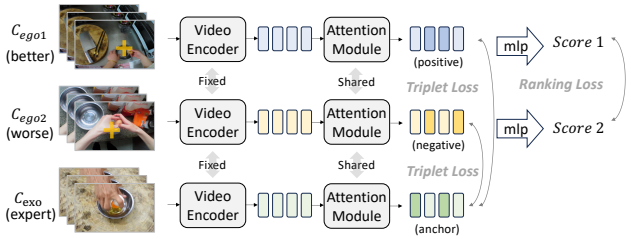
Cross-view action planning. Cross-view action planning utilizes coarse-level annotations with a total of 27 classes for training, validation, and testing. We sort all action steps in each video by their start time. Consequently, this task is oriented towards predicting potential sequences of future action starts. After filtering, we obtain 2.1k/0.8k/1.2k action steps in the egocentric train/validation/test set and 2.4k/0.3k/0.4k action steps in the egocentric train/validation/test set. Note that it is also possible to use the fine-level action annotations for this task, which will result in a much larger dataset split. We do not use this setting since we observe a large variation in the fine-level actions due to practical issues such as environmental constraints and unskilled performance. We believe the combination of our cross-view anticipation and cross-view planning can well evaluate the ability to bridge ego-exo procedural activities at both clip-level and task-level.

Method	Gaze	Anticipation↑				Planning↓	
		Ego-V	Ego-N	Exo-V	Exo-N	Ego	Exo
Exo-only	✗	30.7	23.5	**40.9**	**42.5**	83.5	**74.6**
Ego-only	✗	33.4	37.6	28.7	18.0	82.3	83.7
Ego-only	✓	**40.9**	**52.3**	37.5	37.6	**79.0**	81.8
Ego-only	Center	33.4	38.8	33.1	33.7	81.2	84.4
Unsupervised Domain Adaption							
Ego2Exo	✗	34.1	38.0	34.2	28.4	82.1	83.5
Ego2Exo	✓	**41.0**	**53.7**	37.2	37.3	**81.5**	83.8
Exo2Ego	✗	31.6	24.2	39.9	**42.4**	82.9	77.4
Exo2Ego	✓	34.1	31.5	**40.2**	42.3	81.8	**76.9**
Knowledge Distillation							
Ego2Exo	✗	30.7	25.1	**41.5**	**47.6**	83.0	75.1
Ego2Exo	✓	30.6	25.3	41.0	47.1	83.1	**74.6**
Exo2Ego	✗	34.6	38.3	30.1	18.9	81.9	84.9
Exo2Ego	✓	**41.2**	**55.9**	37.0	39.8	**79.0**	82.6
Co-training							
Ego & Exo	✗	33.9	37.3	**40.3**	46.7	82.0	74.8
Ego & Exo	✓	**41.6**	**52.9**	39.6	**47.9**	**78.3**	**74.4**

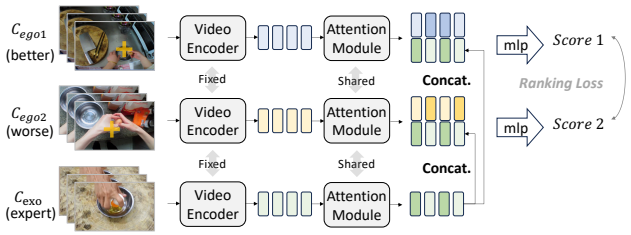
Table S6. Results of cross-view action anticipation and planning benchmarks on the validation set. For anticipation, the class-mean Top-5 recall is used as the evaluation metric (higher is better). For planning, the Edit distance is used as the evaluation metric (lower is better). Gray cells show the cross-view performance.

S6.2.4 Additional results

Table S6 presents the results of our baseline models on the validation set for the cross-view action anticipation and planning bench-



(a) Cross-view skill assessment with triplet loss



(b) Cross-view skill assessment with relation network

Figure S8. Cross-view referenced skill assessment with triplet loss and relation network.

marks. In the first block, zero-shot cross-view evaluation (*e.g., Exo-only* evaluated on ego view, and *Ego-only* evaluated on exo view) results in the lowest performance levels. This outcome underscores the challenge of applying learned representations from one perspective directly to another without any intermediary processing or adaptation. A significant improvement in zero-shot cross-view performance is observed with the introduction of gaze-cropped inputs. This enhancement suggests that **gaze can be an effective bridge for the ego and exo actions**. Further improvements in performance are noted when implementing methods such as Unsupervised Domain Adaptation (UDA), Knowledge Distillation (KD), and Co-Training (CT). The results also demonstrate that the extent of performance improvement varies across different cross-view settings. This variation highlights the complexity of bridging activities in ego and exo views and the importance of selecting the most appropriate method based on the specific requirements of each task.

S6.3. Cross-view referenced skill assessment

S6.3.1 Detailed task definition

Our training dataset comprises the following components: (1) Egocentric Video Pairs: Denoted as P, each pair $(C_{ego1}, C_{ego2}) \in P$ is arranged such that video C_{ego1} displays better skill than C_{ego2}. (2) Accompanying Gaze Sequences: For every pair of egocentric videos $(C_{ego1}, C_{ego2}) \in P$, corresponding gaze sequences (g_1, g_2) are provided. (3) Exo-View Expert Demonstration: Each pair $(C_{ego1}, C_{ego2}) \in P$ is accompanied by an expert demonstration video C_{exo}, showcasing the same action as in C_{ego1} and C_{ego2} from an exo-view perspective. The objective is to develop a ranking function $f(\cdot)$ that adheres to the condition $f(C_{ego1}) > f(C_{ego2})$ given (g_1, g_2) and C_{exo} as the reference.

	#video clip	#valid pairs	Av. length	Corr. exo	Gaze
EPIC-Skills [26]	216	2592	85s	✗	✗
BEST [27]	500	16782	**180s**	✗	✗
Infant Grasp [79]	94	3318	5s	✗	✗
Ours	**3304**	**34239**	10s	✓	✓

Table S7. Comparison of skill assessment datasets based on human pairwise ranking annotation.

S6.3.2 Implementation details

Network architecture. As shown in Fig. S8, we assume C_{ego1} exhibits a higher skill level compared to C_{ego2}. Built upon a pairwise ranking skill assessment model RAAN [27], we employ different video encoders, including I3D [15] and VideoMAE [126], to extract video features from C_{ego1}, C_{ego2}, and C_{exo}. The features are processed by an attention module as described in [27] and resulting in refined features F_{ego1}, F_{ego2}, and F_{exo}. Then, we apply two different approaches to leverage the reference exo-view demonstration video: 1) Triplet loss (TL). We designate F_{exo} as the *anchor*, F_{ego1} as the *similar item (positive)*, and F_{ego2} as the *dissimilar item (negative)*. Then, we apply a triplet margin loss with margin = 1, to aid the model in understanding that the anchor is closer to the positive than the negative item. In our scenario, C_{ego1} demonstrates a skill level closer to the expert. 2) Relation network (RN). Inspired by [124], we implement a relation network that concatenates the features of the ego and exo clips. Precisely, we set $F_{ego1} = Concat(F_{ego1}, F_{exo})$ and $F_{ego2} = Concat(F_{ego2}, F_{exo})$. By combining ego and exo features, this network is designed to implicitly discern which of the two egocentric video clips bears a closer relation to the demonstration video in terms of skill level. Finally, the refined features F_{ego1} and F_{ego2} are processed by an MLP to regress skill scores for the two ego videos.

Training. For the ego branch of our network, we employ the training objectives from [26, 27]. 1) a margin ranking loss is applied on the finally generated scores to ensure $ego1$ is ranked higher than $ego2$. 2) a disparity loss is applied within the attention module to prevent the network from getting trapped in local minima during training 3) a rank-aware loss and a diversity loss are also applied following [27]. Besides the ego branch, to leverage the exo demonstration video, we propose to utilize a triplet loss to aid the model in comprehending that $ego1$ exhibits skills more akin to those of an expert.

S6.3.3 Annotation details

We include two types of annotations for skill level. The first type is self skill assessment. During data collection, subjects are asked to assess themselves on various aspects, including their familiarity with cooking environments, the number of times they have completed the task previously, the frequency of performing the task, the typical duration required to complete the task, and whether they've taught others how to perform the task. Based on the self-evaluation results, we have observed a considerable diversity in subjects' skill levels, which motivates us to craft the skill assessment benchmark. One related work is HoloAssist [137] where they show the distri-

Method	Gaze	Egg Cracking	Peeling	Stir-fry	Cutting	Avg
Ego pairs only						
Who's better* [26]	✗	79.08	74.52	82.87	78.35	78.71
RAAN* [27]	✗	83.09	77.30	86.25	82.86	82.23
Who's better* [26]	✓	79.95	75.67	82.94	79.21	79.44
RAAN* [27]	✓	**84.79**	78.97	86.14	82.96	**83.22**
Ego pairs + Exo						
RAAN* [27] + RN	✗	83.14	77.39	**86.47**	82.48	83.01
RAAN* [27] + TL	✗	81.99	77.48	86.16	82.54	82.04
RAAN* [27] + RN	✓	82.84	78.75	86.19	**83.33**	82.78
RAAN* [27] + TL	✓	83.64	**79.41**	86.14	83.07	83.07

Table S8. Ranking accuracy of cross-view referenced skill assessment. "*" means using VideoMAE [126] extracted video features. In the upper part of the table, only ego video pairs are used, while in the lower part, exo demonstrations are incorporated by "RN": relation network and "TL": triplet loss.

bution of the performers' familiarity with the tasks measured by a self-reported score (0-10) by the subjects. However, no related benchmarks is provided by HoloAssist.

One drawback of self-evaluated skill level is that individuals may showcase varying skill levels in each video instance, even across multiple attempts [26]. As a more objective complement of the self-assessment, we adopt the pairwise comparison approach [26, 27, 79] for annotation. We provide annotators with four criteria: Fluency, Speed, Proficiency, and Skillfulness. These standards serve as the basis for their ranking assessment. From the annotation results, we find 40% of the rankings deviate from the rankings based on the self-evaluations of the two subjects in the video pair. This finding supports that relying solely on self-evaluation is inadequate for creating a robust skill assessment benchmark.

As shown in Table S7, our dataset stands out as the only skill assessment dataset featuring the gaze modality and corresponding exo-view demonstration videos. Notably, our dataset surpasses previous ones in both video clip quantity and valid pair numbers. We follow the setting in [26, 27] to employ 4 individuals to rank the same video pair to ensure credibility. We exclude annotations with fewer than 3 consistent opinions instead of 4 to ensure our dataset contains challenging pairs. Regarding action categories, our dataset comprises 6 actions: Egg cracking, Peeling, Stir-fry, Cutting into chunks, Slicing into strips, and Chopping into pieces. In the main paper, we merge the last three actions into a comprehensive category labeled "Cutting", encompassing various knife-using skills.

S6.3.4 Additional results

Results with I3D [15] feature are shown in Table 5 of the main manuscript. We show the results with VideoMAE [126] feature in Tab. S8. Comparing results from the two tables, we observe an overall increase in performance in all cases in Table S8 because of the stronger backbone model. While we can still observe performance gain when adding Exo reference video, this improvement is less significant compared with the corresponding table in the main manuscript. We suspect that this variation is attributed to the varying degrees of influence that the intrinsic properties of the extracted features exert on the observed enhancements.

S6.4. Cross-view action segmentation

S6.4.1 Detailed task definition

The action segmentation task in our framework is focused on both categorizing each time step and delineating action steps within procedural videos. Given a lengthy video V comprising N_V frames at 25 FPS, the model is tasked with classifying the category of each frame in the video. The evaluation metric includes assessing frame-level classification accuracy. Additionally, sequence-level metrics such as edit distance and instance-level metric F1 are employed for further evaluation [29]. The extended cross-view action segmentation benchmark, similar to cross-view action anticipation and planning, aims to pursue performance improvement by receiving aid from other views.

S6.4.2 Implementation details

Network structure. We employ I3D [15] as the feature extractor to generate temporal features, following the methodology of previous work [29]. To implement our various training settings, we utilize the SSTDA [18] code base. For both training and testing, we downsample feature sequences and label sequences by a factor of 5 for efficiency.

Training. The loss function used to train the action segmentation task is derived from SSTDA [18]. The model consists of multiple stages. The overall loss function for a single stage is a combination of the classification loss and smoothing loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{cls} + \gamma \mathcal{L}_{smooth}. \tag{8}$$

The model is trained using the Adam [60] optimizer with a learning rate set to 1e-3 and the training process spans 150 epochs.

Cross-view settings. Similar to cross-view action anticipation and planning, action segmentation also performs four cross-view settings. Though cross-view action segmentation shows different input and output, which yields dense prediction, the implementation of cross-view settings is consistent with Section S6.2.2.

S6.4.3 Annotation details

The annotation for the cross-view action segmentation task is derived from coarse-level annotations. To create non-overlapping segment annotations for temporal action segmentation, we establish the center point of the overlapping portion of two segments as their boundary. Subsequently, we introduce background segments labeled as "no action" in temporal regions not covered by action annotations. Finally, we obtain 173/57/85 videos in the egocentric train/validation/test set and 210/24/32 videos in the exocentric train/validation/test set.

S6.4.4 Experimental results

Table S9 presents the results of our baseline models on the validation set and test set for the cross-view action segmentation benchmarks. These results mirror the trends observed in the action anticipation and planning benchmarks: without any assistance from another view, the models can only perform well on the test data in

Method	Gaze	Val						Test					
		Ego			Exo			Ego			Exo		
		Acc	Edit	F1@Avg	Acc	Edit	F1@Avg	Acc	Edit	F1@Avg	Acc	Edit	F1@Avg
Exo-only	✗	27.99	33.81	6.95	**38.64**	**40.28**	**23.64**	24.80	35.29	8.13	**42.65**	**37.32**	**20.14**
Ego-only	✗	65.35	44.25	40.91	19.81	21.18	7.31	62.50	44.29	39.43	25.09	22.45	6.89
Ego-only	✓	**66.01**	**46.60**	**41.78**	21.08	23.29	7.44	**65.99**	**48.83**	**42.95**	25.14	22.28	8.17
Ego-only	Center	62.14	45.92	36.60	19.13	23.02	7.37	60.42	46.29	36.52	22.83	22.55	7.2
Unsupervised Domain Adaption													
Ego2Exo	✗	65.52	44.15	40.78	20.78	22.02	7.55	63.41	44.35	40.15	25.67	23.12	7.45
Ego2Exo	✓	**66.12**	**46.42**	**42.11**	21.78	23.99	8.43	**65.91**	**48.81**	**42.78**	25.87	23.44	8.56
Exo2Ego	✗	28.76	33.76	7.56	38.44	39.98	23.61	25.34	35.75	8.67	**42.56**	**36.71**	20.03
Exo2Ego	✓	29.12	34.91	8.49	**38.47**	**40.01**	**23.69**	27.78	39.12	9.87	42.45	36.66	**21.12**
Knowledge Distillation													
Ego2Exo	✗	33.25	25.68	9.18	39.62	40.36	20.07	32.00	26.70	9.73	**43.03**	38.06	**20.44**
Ego2Exo	✓	31.16	28.62	8.60	**40.28**	**42.24**	**23.09**	29.17	28.49	8.45	41.68	**36.33**	20.12
Exo2Ego	✗	65.91	45.80	**42.37**	22.65	17.86	7.16	62.77	46.73	41.12	28.20	17.78	6.06
Exo2Ego	✓	**66.02**	**47.98**	41.71	23.47	24.24	8.12	**64.53**	**49.36**	**42.24**	28.34	23.49	7.80
Co-training													
Ego & Exo	✗	64.43	42.00	37.40	37.93	**42.18**	**23.20**	61.75	41.45	36.43	41.07	**38.73**	21.93
Ego & Exo	✓	**66.57**	**44.36**	**39.87**	**41.89**	39.13	22.70	**65.57**	**44.30**	**39.62**	**42.27**	35.10	**22.50**

Table S9. Results on cross-view temporal action segmentation benchmark. Gray cells show the cross-view performance.

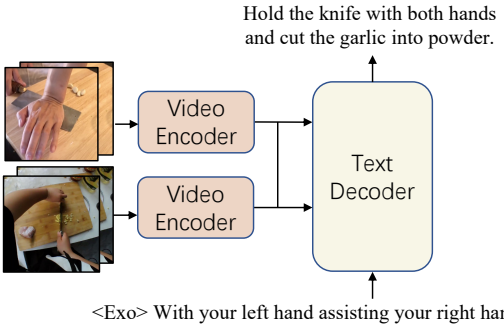


Figure S9. Cross-view referenced captioning with a video encoder and a text decoder.

the same view. The inclusion of gaze data enhances model performance in both the ego-only setting and the cross-view setting. This suggests that focusing on areas of visual attention, as indicated by gaze data, is beneficial for better understanding and segmenting actions, regardless of the viewpoint. When information from another view is leveraged, all three methods – Unsupervised Domain Adaptation (UDA), Knowledge Distillation (KD), and Co-Training (CT) – contribute to performance improvements in the cross-view setting. Each method offers a different mechanism for integrating cross-view insights, thus aiding in the segmentation task. Reflecting the varying degrees of labeled data utilization, Co-Training (CT) tends to outperform Knowledge Distillation (KD), which in turn outperforms Unsupervised Domain Adaptation (UDA).

S6.5. Cross-view referenced video captioning

S6.5.1 Detailed task definition

Cross-view referenced video captioning evaluates the model's captioning ability to leverage cross-view information for caption generation. Our motivation is that egocentric videos require extensive efforts to collect, and are thus limited in scale and diversity. In contrast, large-scale exocentric videos can be easily sourced from the Internet. The question is, *how to leverage such exocentric videos to help the understanding of limited egocentric videos?*.

Formally, at the training stage, we have egocentric videos of limited size $\{(V_1^{\text{ego}}, T_1^{\text{ego}}), ..., (V_N^{\text{ego}}, T_N^{\text{ego}})\}$ with N samples, and exocentric videos $\{(V_1^{\text{exo}}, T_1^{\text{exo}}), ..., (V_M^{\text{exo}}, T_M^{\text{exo}})\}$, where $N \ll M$. Each video is paired with a fine-grained text description. The goal is to train a cross-view video captioning model $f(\cdot)$ using exocentric videos as references. At the inference stage, the model is required to generate the captions of the testing egocentric videos, given the other set of exocentric videos as references. Note that, $N \leq M$ only holds for the training set. In particular, we limit the number of the referenced exocentric videos by formulating the task as a K-shot captioning [2] problem, where K denotes the maximum number of exocentric videos that the model is allowed to use during inference. The inference process can be formulated as $f(V^{\text{ego}}|\{(V_1^{\text{exo}}, T_1^{\text{exo}}), ..., (V_K^{\text{exo}}, T_K^{\text{exo}})\}$ In practice, we consider three settings, 0-shot, 1-shot, and 2-shot.

S6.5.2 Annotation

We directly apply the fine-grained language annotations in our dataset. The referenced exocentric videos are randomly selected for training/validation/testing, respectively. The training set only contains 1000 egocentric videos with 6270 referenced exocentric videos. For the validation/testing set, there are 8181/2143, 18243/4930 egocentric videos and referenced exocentric videos, respectively.

Method	Ref Train	Ref Infer	Validation				Test			
			BLEU-4	METEOR	ROUGE-L	CIDER	BLEU-4	METEOR	ROUGE-L	CIDER
Exo-only	✗	✗	0.024	0.126	0.212	0.122	0.023	0.124	0.208	0.112
Ego-only (0-shot)	✗	✗	0.049	0.116	0.270	0.332	0.048	0.112	0.266	0.314
Co-training										
Ego+Exo	✓	✗	**0.069**	**0.139**	**0.294**	**0.460**	**0.068**	**0.137**	**0.290**	**0.427**
Ref-training										
Ego+Exo (1-shot)	✓	✓	0.047	0.121	0.275	0.378	0.046	0.123	0.275	0.372
Ego+Exo (2-shot)	✓	✓	0.044	0.119	0.272	0.372	0.045	0.122	0.272	0.380

Table S10. Cross-view referenced captioning performance. "Ref Train/Ref Infer" refers to whether the model uses exocentric videos during training/inference.

S6.5.3 Implementation details

For the baseline model, we choose a Flamingo-style captioning model [2, 4, 69], an advanced vision-language model designed for few-shot vision-language tasks, as shown in Fig. S9. Please refer to [2] for the architectural details. We simply pre-pend the referenced video(s) before the input video, and add the referenced caption as prompts to the text decoder. We train the model for 3 epochs using the Adam optimizer, with an initial learning rate of 1e-4 and a batch size of 32. We adopt the cross-view association network (Fig S5(b)) to select referenced samples.

S6.5.4 Results

Table S10 lists the cross-view referenced captioning performance. We consider three baseline models: (i) **Single-view** models include *Ego-only* and *Exo-only*, where the former one merely adopts egocentric videos for training and inference without seeing exocentric videos. The *Exo-only* model uses all referenced exocentric videos for training, and it is then evaluated on egocentric videos. (ii) **Co-training model** is trained on both egocentric videos and referenced exocentric videos, and transferred to egocentric test videos. (iii) **Referenced-training** model refers to our model introduced in Fig. S9, where the model leverages one (1-shot) or two (2-shot) exocentric videos to make predictions. As shown in Table S10, both the co-training model and referenced-training models outperform single-view models. For co-training models, the performance gain is due to the increased number of training data (ego+exo), compared to ego-only and exo-only counterparts. In terms of referenced-training models, they generally outperform the ego-only counterpart by additionally incorporating exocentric videos in the model. Results of both the co-training model and referenced-training models indicate the effectiveness of utilizing exocentric videos in improving egocentric video captioning when the data is of limited scale.

S6.6. Zero-shot action recognition

We assess the zero-shot classification performance of verb and noun subsets. In cases where samples have multiple labels, we straightforwardly replicate the samples for testing. Our testing procedure follows CLIP [107], evaluating the vision-language models based on Top-1 and Top-5 accuracy.

S6.6.1 Annotation

In this task, our evaluation specifically addresses zero-shot transfer within the closed set and does not encompass cross-view settings. It is noteworthy that this annotation does not require ensuring consistent categories between egocentric and exocentric datasets across their respective validation and testing sets. The size of the resulting egocentric verb-validation/verb-test/noun-validation/noun-test set is 14.4k/32.6k/20.2k/44.8k, and the size of the exocentric verb-validation/verb-test/noun-validation/noun-test set is 4.2k/10.4k/5.7k/13.1k, respectively.

S6.6.2 Implementation details

We use the 16 prompts from the zero-shot classification on Kinetics [15] for verb and noun subsets. These prompts are listed in Table S12. We sample the center frame of each video clip, and use OpenAI CLIP [107] to extract the visual features and textual features.

S6.6.3 Experimental results

Table S11 shows the performance of zero-shot action recognition. *Oracle* is the upper bound of accuracy, given that this is a multi-class action recognition problem. On both the validation set and the test set, the zero-shot performance on egocentric videos is worse than that on exocentric videos, particularly in the top-1 accuracy. This result indicates the limitation in cross-view action understanding of the current method.

S6.7. Fine-tuned action recognition

S6.7.1 Detailed task definition

We formulate the conventional Fully-supervised setting to a multi-label classification task. In assessing the performance of fully supervised action recognition, we employ the class-wise multi-label mean Average Precision (Marco mAP) evaluation metric due to the presence of multiple labels per clip. This evaluation protocol is reasonable because it matches the long-tail attribution of actions in `EgoExoLearn`.

S6.7.2 Annotation

In this task, our evaluation focuses on the closed set and does not consider cross-view settings. Thus, our annotations ensure that

Model	Val								Test							
	Ego-Verb		Ego-Noun		Exo-Verb		Exo-Noun		Ego-Verb		Ego-Noun		Exo-Verb		Exo-Noun	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Oracle	56.14	99.79	39.72	99.29	50.06	99.74	36.74	97.70	55.41	99.66	46.55	99.69	45.77	99.72	37.00	97.59
CLIP [107]	7.89	22.71	7.08	19.26	9.49	22.62	7.70	20.45	6.96	21.95	6.39	18.19	9.02	20.99	7.09	19.93

Table S11. Results of zero-shot action recognition. *Oracle* denotes the upper bound of accuracy, because of the multi-label nature of clips in our dataset.

#	Prompts
1	A photo of action {}.
2	A picture of action {}.
3	Human action of {}.
4	{}, an action.
5	{} this is an action.
6	{}, a video of action.
7	Playing action of {}.
8	{}
9	Playing a kind of action, {}.
10	Doing a kind of action, {}.
11	Look, the human is {}.
12	Can you recognize the action of {}?
13	Video classification of {}.
14	A video of {}.
15	The man is {}.
16	The woman is {}.

Table S12. Prompt templates used in the zero-shot action recognition task.

View	Model	Val		Test	
		Verb	Noun	Verb	Noun
Ego	Slowfast-R50 [30] 4×16	27.03	34.77	25.58	33.25
	MViT-S [28]	**29.83**	**39.45**	**28.16**	**36.46**
Exo	Slowfast-R50 [30] 4×16	15.79	22.08	11.71	16.65
	MViT-S [28]	**18.59**	**22.81**	**13.53**	**19.36**

Table S13. Results of fine-tuned action recognition. We utilize the multi-label mean Average Precision (mAP) evaluation metric because of the existence of multiple labels per clip. This choice is consistent with the methodology described in [93]. Specifically, we adopt macro mAP as the class-mean metric.

egocentric and exocentric datasets maintain consistent categories across their respective training, validation, and testing sets. At last, this task contains 81/211 verb/noun categories in the egocentric set and 69/183 verb/noun categories in the exocentric set. The size of the egocentric train/validation/test set is 36k/8k/18k, and the size of the exocentric train/validation/test set is 6.2k/2.1k/4.8k.

config	Egocentric	Exocentric
optimizer	AdamW [60]	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
weight decay	1e-4 (Slowfast), 0.05 (MViT)	
learning rate scheduler	warmup constant	
learning rate	1e-4	
batch size	32	32
total epochs	20	30
flip augmentation	✓	
crop size	224	
randomresizedcrop	scale=(0.08, 1)	

Table S14. Training hyperparameters used for fine-tuned action recognition benchmark.

S6.7.3 Implementation details

For evaluating this task, we utilize SlowFast-R50 [30] and MViT-Small [28] as the backbones. The weights pretrained on the Kinetics [15] dataset are employed for both backbones. Frames within each action clip are uniformly sampled and fed into the backbone. The multi-label classification task is supervised using the standard cross-entropy loss. Table S14 lists the training hyperparameters.

S6.7.4 Experimental settings and results.

Table S13 shows the result of fine-tuned action recognition. MViT-S [28] (with 16 frames input) exhibits superior performance and generalization compared to the R50-based SlowFast [30] (with 4 frames for the slow branch and 32 frames for the fast branch as input). The results in Table S13 also reveal great potential improvement on more sophisticated model structures for this dataset.

S7. Additional Dataset Details

S7.1. Language annotation

Different from previous datasets [22, 37, 137], our dataset includes two-level language annotations with manually annotated temporal boundaries. As described in Section 3.2 of the main manuscript, our annotation includes a coarse-level language annotation and a fine-level language annotation. We designed a web-based interface to facilitate the annotation. An example screenshot is shown in Figure S10.

For each video, the annotators are asked to quickly skim the video to grab the overall content, and then begin the annotation

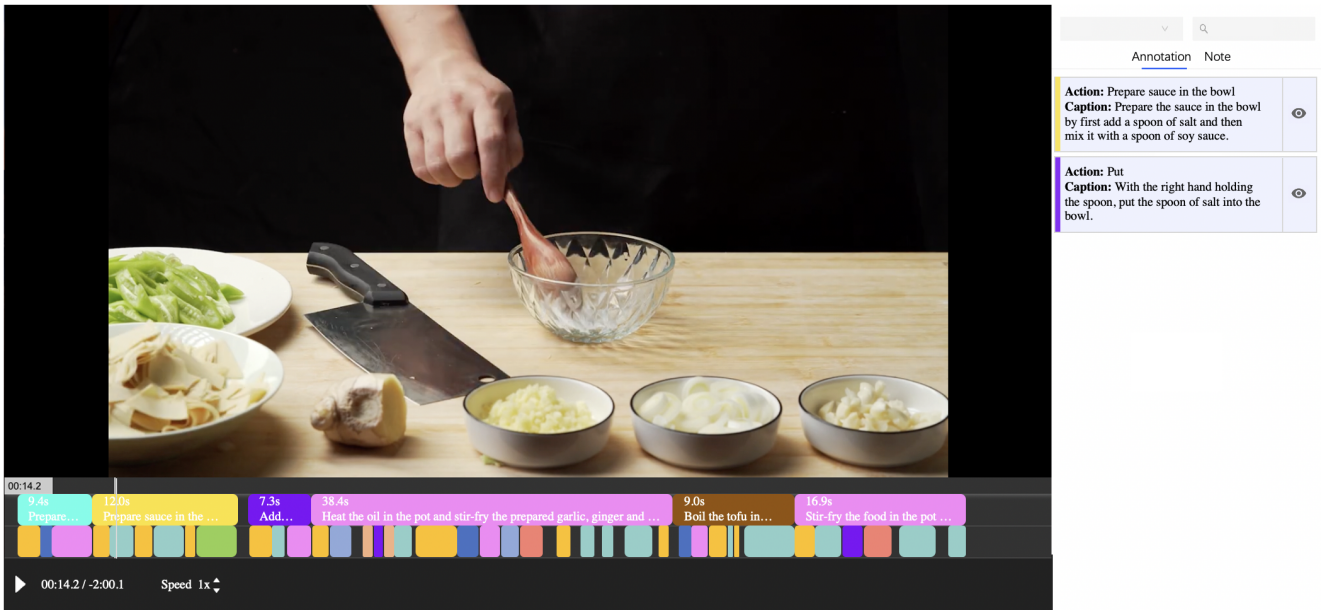


Figure S10. We use a web-based language annotation interface for the annotators. Annotators mark a segment of the video, select a category for this segment, and describe the segment based on the annotation requirement in their mother language.

of each session. For the daily tasks, the annotators are instructed to describe each segment based on their own knowledge. For the tasks in specialized laboratories, we train the annotators showing them the process of the experiments, the technical terms of some tools/reagents (*e.g.*, pipette), and the purpose of each action step. To avoid describing objects that are impossible to determine visually (*e.g.*, the appearance of water and PBS reagent are exactly the same), we ask the annotators to describe their visual appearance instead (*e.g.*, pink reagent in a bottle with green cap). Figure S11 shows a word cloud of the language annotations separated by views and tasks. Figure S13 shows the distribution of lengths of the coarse and fine level language annotations. The average lengths of the coarse and fine level annotations are 21.5 seconds and 4.6 seconds, respectively.

Translation & Parsing. For all the non-English language annotations, we translate them into English using ChatGPT. We conduct a manual check on the translation quality and use Google Translation API to translate again for unsatisfactory translations.

To effectively parse and analyze the annotations in our dataset, we employ a rule-based framework designed to extract verbs and nouns associated with specific actions of the left and/or right hand. The process is methodical and iterative to ensure the annotation quality. The overview of the parsing is as follows:

- Sentence Splitting: We begin by splitting the annotations into individual sentences using separators like commas. This step helps in isolating distinct actions or descriptions for more focused analysis.
- Keyword Identification and Extraction: For each split sentence, we use NLTK to identify keywords that indicate actions related to the left hand, right hand, both hands, etc. This involves analyzing the sentence structure and content to pinpoint relevant verbs and nouns. One challenge we encounter is the word "left" itself, which can be a verb in certain contexts. To address this, we temporarily mask the mentions of left and right hands in each sentence and then re-extract the verbs and nouns. This masking helps in distinguishing between the directional use of "left" and its use as a verb.
- Manual Review and Iteration: After the initial extraction, we conduct a manual review of the results to identify and correct any errors. This step is crucial for ensuring the accuracy and relevance of the extracted terms. If errors are found, we revisit the first and second steps, making necessary adjustments. This iterative process continues until the manual review yields satisfactory results.

Figure S12 shows the verbs and nouns extracted after associating with the left and right hands. We only show the top 30 categories due to the size limit.

S7.2. Post-processing.

In dealing with the practical challenges of recording egocentric videos, particularly with Pupil Invisible devices that sometimes capture footage at variable frame rates due to issues like overheating, we employ post-processing for standardization. All videos are converted to a constant frame rate of 25fps to ensure uniformity and consistency in our dataset.

Additionally, our gaze data, which is recorded at a high frequency of 120Hz, provides detailed insights into the viewer's point of focus during the demonstration following process of the video. To ease the use of this gaze data, we align the timestamps of the egocentric camera with the eye-tracker. Once the alignment process is complete, we register each gaze data point to the temporally closest frame in the video. We then take the average of all gaze data points within one frame and use this as the final gaze data.

In Figure S14, we visualize the video frames along with the

Figure S11. Word cloud of annotations separated by views and tasks.

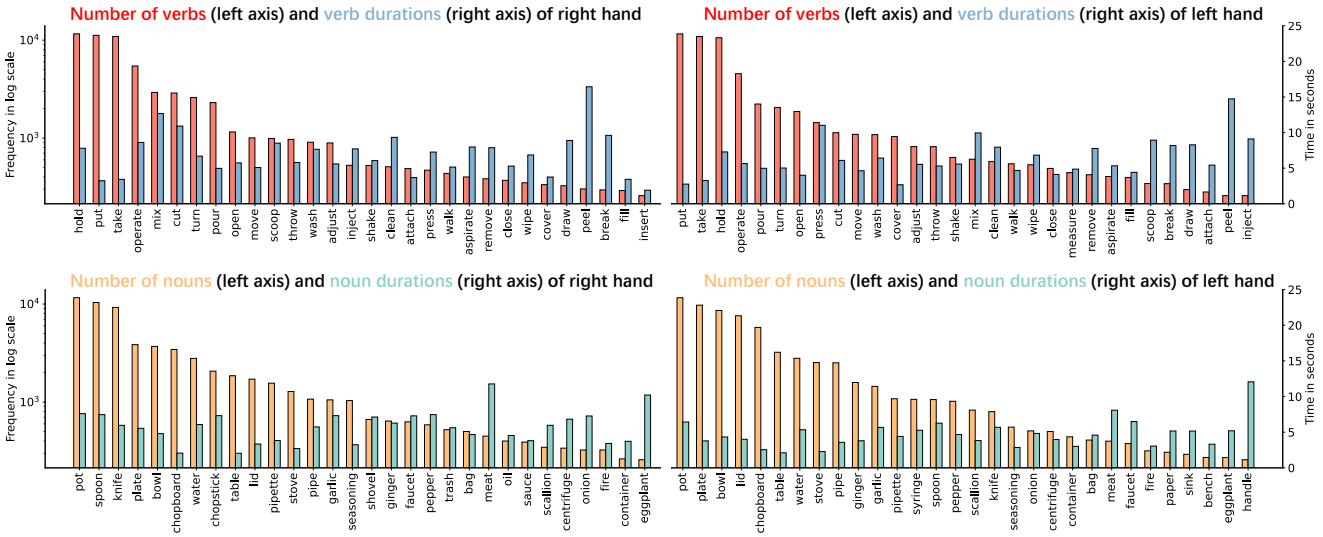


Figure S12. Occurrence and duration distribution of the annotated fine-level verbs and nouns associated with the left and right hands.

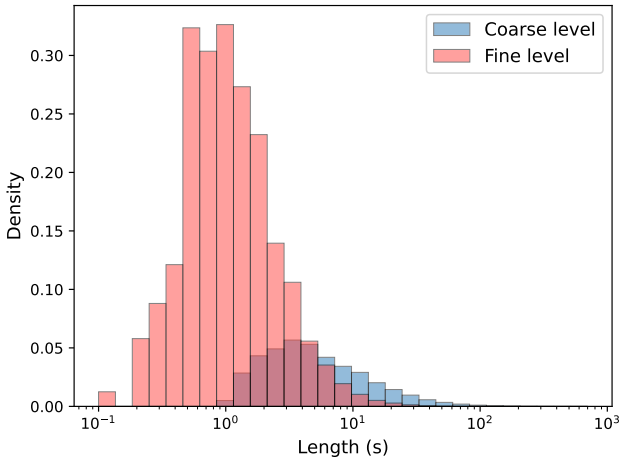


Figure S13. Distribution of the lengths of the coarse-level and fine-level language annotations.

annotated fine-level language annotations. EgoExoLearn features a new demonstration following setting that is a complement to existing egocentric and ego-exo datasets. Meanwhile, as can be seen in Figure S14, compared with existing egocentric datasets, our language annotations contain much longer sentences, enabling our dataset to be used in the captioning benchmarks.

S7.3. IRB approval

We receive IRB approval before the data collection, adhering to the ethical standards and guidelines for research involving human participants. Participants involved in the study were provided with detailed consent forms and information sheets. These documents thoroughly explained the data capture process, the purpose of the study, and how the data would be used in the future. The consent forms, along with the information sheets, were reviewed and approved by the IRB to ensure they met all ethical standards and adequately informed participants. We maintain these documents and can provide them upon request for verification or further inquiry into our ethical and procedural practices during the data collection.

S7.4. Tasks

Our dataset is collected for 5 types of daily tasks and three types of specialized laboratory tasks. The collection is performed in four different kitchens and three different specialized laboratories. The

Figure S14. Examples of video frames and corresponding fine-level language annotations in our dataset.

participants' ages range from 18 to 40 years with diverse occupations such as athletes, housekeepers, security guards, university students, and researchers. We carefully choose the daily tasks and specialized lab tasks such that a long series of procedures is needed before finishing. This can reflect the complexity of real-life activities meanwhile enabling our new benchmarks for ego-exo procedural activity bridging. Table S15 shows the names of the 8 tasks with an example procedure. In real recordings, the procedures are usually more complicated due to repetition and other practical issues related to the environment. Note that the scientific name of the specialized reagents are not included in the language annotations but are described using their visual appearance.

Acknowledgements. Hosted by Shanghai AI Laboratory, Shenzhen Institutes of Advanced Technology, and Nanjing University, this work is jointly completed by a team of researchers and students from multiple institutes. Authors can be contacted via email: Yifei Huang, Mingfang Zhang, and Lijin Yang at [hyf, mfzhang, yang-lj]@iis.u-tokyo.ac.jp, Guo Chen at 602022330002@smail.nju.edu.cn. Jilan Xu can be contacted via 18210240039@fudan.edu.cn, Baoqi Pei via 12321251@zju.edu.cn, Hongjie Zhang via nju.zhanghongjie@gmail.com, and Lu Dong via dl1111@mail.ustc.edu.cn. We give special appreciations to Yi Liu (yi.liu1@siat.ac.cn) for providing insightful ideas in the annotation phase, and Ruijie Zhang (zhangruijie@pjlab.org.cn) for the help in the arrangement of annotators.

Task name	Scenario	Example procedure
Task1: Twice-cooked Pork	Daily	1. Prepare spices: Take out and cut some scallion, ginger, and garlic for later use. 2. Prepare the pork: boil the pork together with scallion and ginger to remove impurities. Drain and set aside. Wash the pot if necessary. 3. Prepare vegetables: Take some pepper and onion, wash, and discard unused parts. 4. Cut vegetables: Cut the prepared vegetable into slices, and put the slices into a plate for later use. 5. Cut the pork: Remove the water on the pork. Use a knife to cut the pork into slices, and set aside for later use. 6. Stir-fry the pork: Add oil into heated pot. Then add the prepared spices into the pot. After stir-frying for about 10 seconds, add the pork into the pot. 7. Stir-fry the vegetables: Heat the pot and add oil, then add vegetables into the pot. Stir-fry until the vegetable is well-cooked, then add the pork into the pot. 8. Add seasoning: Add salt, soy sauce, sugar into the pot. Stir-fry a few times to evenly distribute the flavors. 9. Transfer: Transfer the cooked Twice-cooked Pork from the pot to a plate. Wash the pot if necessary.
Task2: Tofu Skin with Hot Pepper	Daily	1. Prepare tofu skin: Take out the tofu skin, fold them and cut the tofu skin into slices. 2. Prepare hot pepper: Take out some hot pepper, squeeze by hand and then cut into pieces. 3. Prepare spice: Take out and cut some scallion, ginger, and garlic for later use. 4. Boil tofu skin: Put some water into the pot, add baking soda. Boil the tofu skin until the water becomes cloudy. 5. Wash tofu skin: Take out the tofu skin and put them into cold water. Wash the tofu skin such that the smell of soda diminishes. Take out and drain water. 6. Prepare sauce in the pot: Heat up the pot and add some oil. Put the spices into the pot. Use a spoon to put some water, soy sauce, and salt into the pot and heat up until the water boils. 7. Cook tofu skin: Put the tofu skin into the pot, and continue to boil until the pot becomes dry. 8. Cook hot pepper: Add hot pepper into the pot, stir-fry for several times. 9. Transfer: Add some oil into the pot, then transfer the cooked dish into a plate.
Task3: Stir-fried potato, eggplant and green pepper	Daily	1. Prepare potato: Take out some potatoes, peel and clean them. 2. Prepare eggplant: Take out some eggplants, remove the stems, and clean them. 3. Prepare green pepper: Take out some green pepper, remove the stems, and clean them. 4. Cut green pepper: Squeeze the green pepper using the side of the knife, then cut them into pieces. 5. Cut eggplant: Rolling cut the eggplant into pieces. Use hand to squeeze water out of the eggplant pieces. Put some cornstarch onto the eggplant pieces and mix well. 6. Cut potato: Cut the potatoes into pieces. 7. Prepare spices: Take out and cut some scallion, ginger, and garlic for later use. 8. Prepare sauce: Take out a bowl. Add water, soy sauce, cornstarch, salt, sugar, vinegar, cooking wine into the bowl and mix them. 9. Boil potatoes: Boil some water and put the potatoes in. Take the potatoes out when the edges become transparent. 10. Fry vegetables: Add oil into the pot, heat the oil up and fry the peppers first and then the eggplants and then the potatoes. 11. Stir-fry: Add some oil into the pot, heat up and put the spices into the pot. Stir-fry a few times. Add the prepared sauce into the pot and then add all the vegetables. Stir-fry until the vegetables and the sauce are well mixed. 12. Transfer: Transfer the cooked dish from the pot into a plate.
Task4: Moo Shu Pork	Daily	1. Cut pork: Take out a piece of pork and cut into small slices. 2. Prepare pork: Put the pork into a bowl. Add some water and wash. Squeeze the pork and pour the water. Put the pork back and add oil, salt, and cooking wine. Mix well. 3. Prepare vegetables: Wash the necessary vegetables, use the pot to boil the vegetables. Take out for later use. 4. Prepare egg: Crack some eggs into a bowl, mix the eggs. 5. Boil vegetables: Boil some water in the pot. Add vegetables and continue to boil for a minute. 6. Fry eggs: Add oil into the pot and then fry the mixed egg. Put the fried egg scramble into a bowl. 7. Stir-fry vegetables: Heat the pot and add oil. After the oil gets heated, first add the prepared spices and then add the vegetables. Stir-fry the vegetables. 8. Stir-fry pork: Without taking the vegetables out of the pot, add pork into the pot, stir-fry all ingredients together. 9. Stir-fry egg: Without taking the ingredients out of the pot, add scrambled egg into the pot, stir-fry all ingredients together. 10. Transfer: Transfer the cooked dish from the pot into a plate.

Task5: Tomato dough drop soup	Daily	1. Prepare spices: Take out and cut some scallion and cumin for later use. 2. Prepare tomatoes: Take out tomatoes, wash and peel. 3. Cut tomatoes: Use a knife to cut the tomatoes first into slices and then into small pieces. 4. Prepare eggs: Crack eggs into a bowl, then stir until evenly mixed. 5. Fry eggs: Heat oil in a pan, add the evenly mixed egg mixture, and stir-fry, finally transfer the cooked scrambled eggs to a plate. 6. Stir-fry tomatoes: Put the chopped tomatoes into the pan, stir-fry them, and then add the scrambled eggs. 7. Soup-making: Add a large amount of clear water to the pot and bring it to a boil. 8. Prepare dough: Gradually add water to the flour while stirring until the flour forms dough. 9. Soup-making: Drop the flour dough into the boiling soup, while adding them, stir continuously. 10. Add seasoning: Add salt, pepper, and MSG (if desired) to the soup. 11. Transfer: Transfer the soup into a large bowl.
Task6: Solid Phase Peptide Synthesis	Chemical lab	1. Weighing: Use a balance to weigh the desired amount of amino acid powder (white powder). Put the powder into a test tube. 2. Reaction: Use a pipette to aspirate some SPPS resin (Transparent liquid) into the test tube. Shake the test tube and put the test tube onto the shaker machine. 3. Deprotection: Add the needed reagent into the tube to separate resin and peptide. 4. Suction Filteration: Take the test tube from the shaker machine, wash the peptide inside the tube, and suck the liquid into the vacuum tube. 5. Checking: Manual check and take necessary notes.
Task7: Total Protein Extraction	Medical lab	1. Preparation: Take out several test tubes, add the necessary amount of PBS reagent (transparent liquid). Take out the cells from the fridge, disinfect, and warm the cells. 2. Wash cells: Use a pipette to transfer the cells into test tubes. 3. Centrifuge: Balance the test tubes in the centrifuge and then start the centrifugation. 4. Reagent making: Prepare some petri dishes, mark each dish, and add the complete medium (pink liquid) into each dish. 5. Transfer cells: Take the cells out of the centrifuge, and check the cell state. Transfer the cells into the prepared petri dish. 6. Quantification: Use an electron microscope to check the cells and record the required information. 7. Other necessary steps: Repeat necessary steps, make necessary reagents, etc.
Task8: Cell subculture	Biology lab	1. Preparation: Prepare the cell, test tubes, reagents, and petri dishes. Mark accordingly. 2. Wash Cells: Use a pipette to aspirate PBS reagent, use PBS to wash the cells. 3. Digestion: Use a separate pipette to add pancreatic enzymes (pink liquid) into the cells, put the cells into an incubator and wait for 3 minutes. 4. Quantification: Use an electron microscope to check the cells and record the required information. 5. reagent making: Prepare some petri dishes, mark each dish, and add the complete medium (pink liquid) into each dish. Use the complete medium to wash the petri dish. 6. Centrifuge: Balance the test tubes in the centrifuge and then start the centrifugation. 7. Transfer cells: Take the cells out of the centrifuge, and check the cell state. Transfer the cells into the prepared petri dish. 8. Incubation: Put the cells with the petri dish into the incubator.

Table S15. The tasks in our EgoExoLearn with example procedures.

References

[1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017. 8

[2] 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*, 35:23716–23736, 2022. 6, 7

[3] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018. 6

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 7

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3, 6

[6] Albert Bandura. Observational learning. *The international encyclopedia of communication*, 2008. 1, 4

[7] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[8] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[9] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 8

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 5, 1

[11] Vinay Bettadapura, Irfan Essa, and Caroline Pantofaru. Egocentric field-of-view localization using first-person point-of-view devices. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2015. 3

[12] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL Interactive Presentation Sessions*, 2006. 4

[13] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. 1

[14] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016. 3

[15] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 5, 7, 8

[16] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 3

[17] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2, 7

[18] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[19] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[20] Victor G Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2022. 7

[21] 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

[22] 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)*, 130:33–55, 2022. 1, 2, 4, 5, 7, 8

[23] 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. 2

[24] 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. 2

[25] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador,

Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[26] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8, 4, 5

[27] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8, 4, 5

[28] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021. 8

[29] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[30] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 8

[31] Mitch J Fryling, Cristin Johnston, and Linda J Hayes. Understanding observational learning: An interbehavioral approach. *The Analysis of verbal behavior*, 27:191, 2011. 1

[32] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2010. 7

[33] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 3

[34] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[35] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 7

[36] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[37] Kristen Grauman, Andrew Westbury, and Eugene Byrne et al. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 5, 6, 7, 8

[38] Kristen Grauman, Andrew Westbury, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 2

[39] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 7

[40] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 6

[41] Nicola J Hodges, A Mark Williams, Spencer J Hayes, and Gavin Breslin. What is modelled during observational learning? *Journal of sports sciences*, 25(5):531–545, 2007. 1, 4

[42] Jiang Hua, Liangcai Zeng, Gongfa Li, and Zhaojie Ju. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, 21(4):1278, 2021. 1

[43] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6:1049, 2015. 3

[44] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[45] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

[46] Yifei Huang, Minjie Cai, and Yoichi Sato. An ego-vision system for discovering human joint attention. *IEEE Transactions on Human-Machine Systems*, 50(4):306–316, 2020. 3

[47] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[48] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[49] Thomas E Hutchinson, K Preston White, Worthy N Martin, Kelly C Reichert, and Lisa A Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on systems, man, and cybernetics*, 19(6):1527–1534, 1989. 3

[50] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Proceedings of the Conference on Robot Learning*, 2022. 1

[51] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2019. 3

[52] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[53] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos.

In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[54] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[55] Soo-Han Kang and Ji-Hyeong Han. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics*, pages 1–11, 2021. 3

[56] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014. 3

[57] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 3

[58] Daekyum Kim, Brian Byunghyun Kang, Kyu Bum Kim, Hyungmin Choi, Jeesoo Ha, Kyu-Jin Cho, and Sungho Jo. Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. *Science Robotics*, 4(26), 2019. 3

[59] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[60] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 8

[61] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2

[62] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 7

[63] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 5

[64] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision (IJCV)*, pages 1–18, 2023. 3

[65] Stéphane Lallée, Eiichi Yoshida, Anthony Mallet, Francesco Nori, Lorenzo Natale, Giorgio Metta, Felix Warneken, and Peter Ford Dominey. Human-robot cooperation based on interaction learning. *From motor learning to interaction learning in robots*, pages 491–536, 2010. 1

[66] Kyle Lam, Junhong Chen, Zeyu Wang, Fahad M Iqbal, Ara Darzi, Benny Lo, Sanjay Purkayastha, and James M Kinross. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ digital medicine*, 5(1):24, 2022. 8

[67] Jangwon Lee and Michael S Ryoo. Learning robot activities from first-person human videos using convolutional future regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 1

[68] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[69] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 7

[70] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021. 2

[71] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Proceedings of the Conference on Robot Learning*, 2023. 5

[72] Haoxin Li, Yijun Cai, and Wei-Shi Zheng. Deep dual relation modeling for egocentric interaction recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[73] Jiayi Li, Tao Lu, Xiaoge Cao, Yinghao Cai, and Shuo Wang. Meta-imitation learning by watching video demonstrations. In *Proceedings of the International Conference on Learning Representations*, 2021. 1

[74] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 3

[75] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[76] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3, 5

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

[78] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6, 7

[79] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. Manipulation-skill assessment from videos with spatial attention network. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2019. 8, 4, 5

[80] Zhenqiang Li, Lin Gu, Weimin Wang, Ryosuke Nakamura, and Yoichi Sato. Surgical skill assessment via video semantic aggregation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022. 8

[81] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6, 1

[82] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[83] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[84] Xiaotian Liu, Hector Palacios, and Christian Muise. Egocentric planning for scalable embodied task achievement. *arXiv preprint arXiv:2306.01295*, 2023. 1

[85] Yan Liu, Pei Yun Hsueh, Jennifer Lai, Mirweis Sangin, MarcAntoine Nüssli, and Pierre Dillenbourg. Who is the expert? analyzing gaze data to predict expertise level in collaborative applications. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, 2009. 8

[86] Yueyue Liu, Zhijun Li, Huaping Liu, and Zhen Kan. Skill transfer learning for autonomous robots and human–robot cooperation: A survey. *Robotics and Autonomous Systems*, 128:103515, 2020. 1

[87] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[88] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1

[89] Javier Marina-Miranda and V Javier Traver. Head and eye egocentric gesture recognition for human-robot interaction using eyewear cameras. *IEEE Robotics and Automation Letters*, 7(3):7067–7074, 2022. 3

[90] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2019. 2

[91] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1, 6

[92] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1069–1078, 2021. 3

[93] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9434–9445, 2021. 8

[94] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[95] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[96] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[97] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[98] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[99] Boxiao Pan, Bokui Shen, Davis Rempe, Despoina Paschalidou, Kaichun Mo, Yanchao Yang, and Leonidas J Guibas. Copilot: Human-environment collision prediction and localization from egocentric videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1

[100] Hyung Min Park, Seok Han Lee, and Jong Soo Choi. Wearable augmented reality system using gaze interaction. In *Proceedings of the IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2008. 3

[101] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[102] Leo Pauly, Wisdom C Agboh, David C Hogg, and Raul Fuentes. O2a: one-shot observational learning with action vectors. *Frontiers in Robotics and AI*, 2021. 1

[103] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the*

Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 3

[104] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023. 1, 3

[105] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978. 1

[106] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[107] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 5, 7, 1, 2, 8

[108] 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 Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2, 5

[109] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5

[110] Richard Ramsey, David M Kaplan, and Emily S Cross. Watch and learn: the cognitive neuroscience of learning from others' actions. *Trends in Neurosciences*, 44(6):478–491, 2021. 1

[111] Yosef Razin and Karen Feigh. Learning to predict intent from gaze during robotic hand-eye coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 3

[112] Radiah Rivu, Yasmeen Abdrabou, Ken Pfeuffer, Augusto Esteves, Stefanie Meitner, and Florian Alt. Stare: gaze-assisted face-to-face communication in augmented reality. In *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA)*, 2020. 3

[113] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004. 1

[114] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7

[115] Stefan Schaal. Learning from demonstration. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 1996. 1

[116] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 5, 7

[117] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[118] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[119] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 1, 2, 5

[120] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2

[121] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[122] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[123] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[124] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8, 4

[125] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[126] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4, 5

[127] Yuli Vasiliev. *Natural language processing with Python and spaCy: A practical introduction*. 2020. 4

[128] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction.

In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[129] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(3):1–24, 2021. 5

[130] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 1

[131] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[132] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[133] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[134] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[135] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 6, 7

[136] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[137] 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 International Conference on Computer Vision (ICCV)*, 2023. 1, 4, 5, 8

[138] Xiaohan Wang, Yuehu Liu, Xinhang Song, Beibei Wang, and Shuqiang Jiang. Generating explanations for embodied action decision from visual observation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 5

[139] Yeping Wang, Gopika Ajaykumar, and Chien-Ming Huang. See what i see: Enabling user-centric robotic assistance using first-person demonstrations. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 639–648, 2020. 1

[140] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6

[141] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 6

[142] Pengfei Wei, Lingdong Kong, Xinghua Qu, Yi Ren, Jing Jiang, Xiang Yin, et al. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 7

[143] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 3

[144] Mark Wilson, John Mcgrath, Samuel Vine, James Brewer, David Defriend, and Richard Masters. Psychomotor control in a virtual laparoscopic surgery training environment: gaze control parameters differentiate novices from experts. *Surgical Endoscopy*, 24(10):2458–2464, 2010. 8

[145] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[146] Xinxiao Wu, Han Wang, Cuiwei Liu, and Yunde Jia. Cross-view action recognition over heterogeneous feature spaces. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 6

[147] Haifeng Xia, Pu Wang, and Zhengming Ding. Incomplete multi-view domain adaptation via channel enhancement and knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[148] Jianjia Xin, Lichun Wang, Kai Xu, Chao Yang, and Baocai Yin. Learning interaction regions and motion trajectories simultaneously from egocentric demonstration videos. *IEEE Robotics and Automation Letters*, 2023. 3

[149] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 6

[150] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[151] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. *arXiv preprint arXiv:2401.00789*, 2024. 3

[152] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 6

[153] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo.

Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[154] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 6, 7

[155] Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato. Finebio: A fine-grained video dataset of biological experiments with hierarchical annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 2

[156] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3333–3343, 2022. 1

[157] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[158] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024. 5

[159] Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 2

[160] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018. 1

[161] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 8

[162] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2

[163] Thorsten O Zander, Matti Gaertner, Christian Kothe, and Roman Vilimek. Combining eye gaze input with a brain–computer interface for touchless human–computer interaction. *International Journal of Human–Computer Interaction*, 27(1):38–51, 2010. 3

[164] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[165] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[166] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[167] Weiyu Zhang, Menglong Zhu, and KG Derpanis. From actemes to action: A strongly supervised representation for detailed action understanding. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 2

[168] Zehua Zhang, David Crandall, Michael Proulx, Sachin Talathi, and Abhishek Sharma. Can gaze inform egocentric action recognition? In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, 2022. 3

[169] Zhenning Zhang, Zhigeng Pan, Weiqing Li, and Zhiyong Su. X-board: an egocentric adaptive ar assistant for perception in indoor environments. *Virtual Reality*, 27(2):1327–1343, 2023. 5

[170] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 1

[171] Jingjing Zheng and Zhuolin Jiang. Learning view-invariant sparse representations for cross-view action recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 7

[172] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3