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 [4, 34, 100, 105]. With this ability to asynchronously bridge activities from egocentric and exocentric views [95, 103], 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 [24], *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 [94]. However, different from humans, training these AI agents usually requires demonstration videos taken in a similar environment [75, 119] and from a congruent perspective with the AI agents, (*e.g.*, the egocentric point of view [43, 59, 108, 134]). While great effort has been made into the collection of egocentric data in different scenarios [18, 30, 106], it remains crucial for the AI agents to directly learn from demonstration videos taken in a different place and from a different viewpoint [35, 148]. Realizing this capability can unleash the full potential of public instructional video data [82] and is also useful in the human-robot cooperation scenario, especially in novel environments [57, 77, 128].

Current works towards this goal can be roughly divided into two directions. One way is to learn models in simulated environments [11, 64, 79, 89, 92], but it remains difficult for models in this setting to generalize in the real world [126]. 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 [124, 143]. Meanwhile, the few existing datasets in this direction [99, 106, 109] 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.

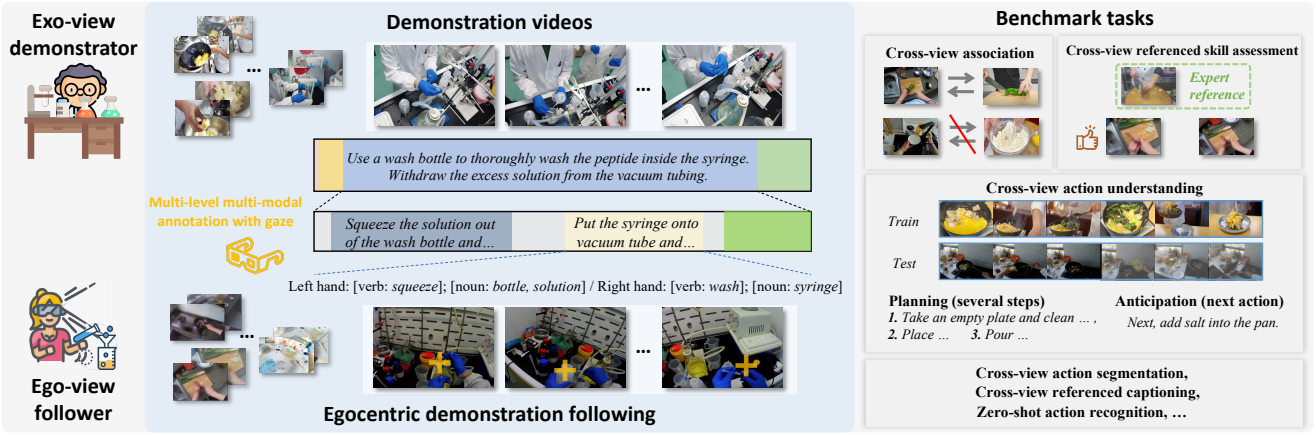

Figure 1. EgoExoLearn emulates the human asynchronous demonstration following process. It contains demonstration videos of multiple tasks, together with egocentric videos recorded by participants replicating the procedure after watching the demonstrations. The dataset comprises gaze signals and fine-grained 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 [14, 144, 147] or quality [124], meanwhile focusing only on single actions captured from the same view [81, 88, 131]. As for actions from different views, apart from multi-view fixed camera datasets [6, 15, 54, 61], there also exist datasets with both ego- exo-centric view videos [30, 31, 99, 106, 109]. These datasets are either recorded in the same environment [45, 99, 109] or record time-synced multi-view videos in the same environment with primary focuses on pose/activity understanding grounded in the 3D world [55, 106, 154]. 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 [143], 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 [18, 20, 53, 55, 155] 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 [110], multiple egocentric video datasets [5, 17, 19, 30, 46, 78, 98, 112, 150] 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 [36, 156], thus is used in a diverse range of areas such as human-computer interaction [42, 51, 151], and augmented reality [90, 102]. In computer vision, efforts have been made to leverage gaze in various tasks [37–39, 56, 65–67, 83, 101, 139]. However, with the previous absence of large-scale egocentric datasets that include gaze, this avenue of research is currently underexplored [44, 68, 94, 154, 160]. 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 [13, 26, 27, 40, 50, 93, 96, 113, 120, 121, 125], hand detection [12, 29, 107, 153], and video-language understanding [41, 48]. These form fundamental building block techniques of embodied AI [85], VR/AR [9, 47, 74, 123], and human-robot interaction [63, 80, 86, 137]. Since most egocentric datasets are smaller in scale compared with general datasets which contain mostly exocentric view videos [3, 142, 152], it is possible to leverage exocentric video data to improve model performance on egocentric videos [132]. There are typically three main directions: joint view-invariant learning [124, 140, 142], domain adaptation [136], and knowledge distillation [69]. 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 [49], 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.

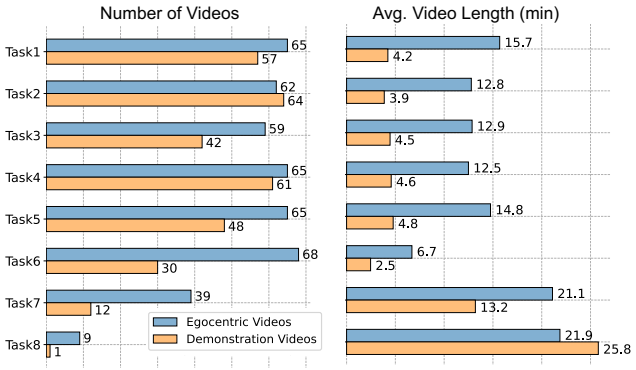

Figure 2. The number of videos per task (left) and the average duration of each video per task (right). Task1 to Task5 represent the 5 daily tasks and the remaining are three tasks in specialized laboratories. In each recording session of the egocentric video, one participant may learn from multiple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

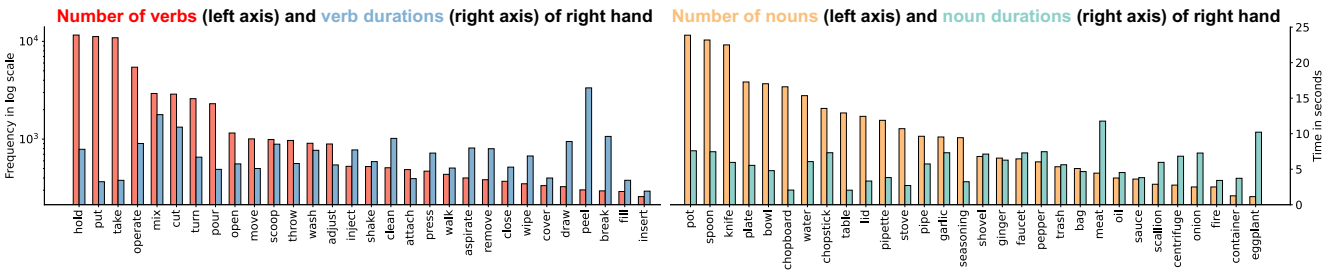

Figure 3. Occurrence and duration distribution of the annotated fine-level verbs and nouns associated with the right hand.

the previous works [106, 126], 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 [18]. 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 [10] and Spacy [116] 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 [30], 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 [4, 34] 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 [98]	Industry	7	✗	✓	✗	✗	✗	✓	✗	✗	✗
EGTEA [67]	Cooking	28	✗	✓	✗	✓	✗	✓	✗	✗	✗
EK-100 [18]	Cooking	100	✗	✗	✗	✗	✗	✓	✗	✗	✗
HoloAssist [126]	Assistive	166	✗	✓	✗	✓	✓	✓	✗	✗	✓
Ego4D [30]	Multiple	3670	✗	✗	✗	◇	◇	◇	◇	✗	✗
H2O [55]	Desk	1	✓	✗	✗	✗	✗	✓	✗	✗	✗
LEMMA [45]	Daily	10	✓	✗	✗	✗	✗	✓	✗	✗	✗
HOMEAGE [99]	Daily	25	✓	✗	✗	✗	✓	✓	✗	✗	✗
CharadesEgo [109]	Daily	34	✓	✗	✗	✗	✗	✓	✗	✗	✗
Assembly101 [55]	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 [109]	7860	0.5	69k	1.8	32	37	1	1	0
HOMEAGE [99]	5700	0.9	26k	-	29	86	-	-	0
EK-100 [18]	700	8.5	90k	3.1	97	300	1	1.2	3.0
Assembly101 [106]	4321	7.1	83k	1.7	24	90	1	1	0
Ego4D [30]	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 [30], we select the closest subtask of "forecasting" following [106].

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 [21, 62].

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 [18, 126]. 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 [157]. Another potential application is the embodied AI agent that can explain its decision [118, 127, 146].

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 [97]. We use TimeSformer-B [8] as the video encoder and clip-text [97] as the text encoder. We initialize both encoders

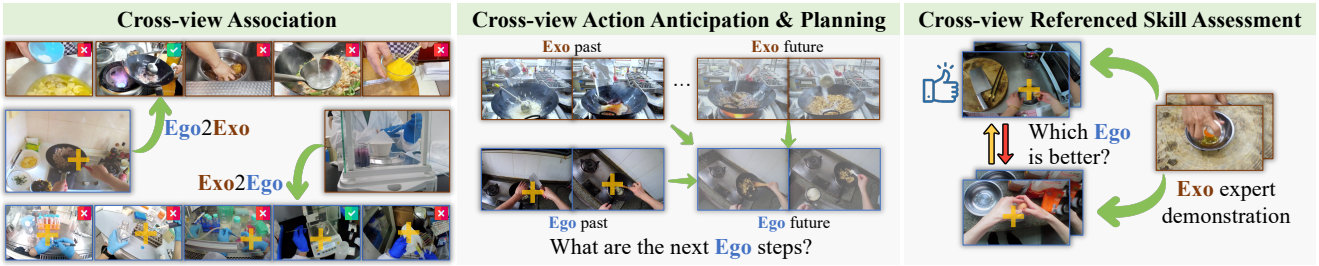

Figure 4. Concept of the 3 benchmarks of cross-view association (Sec. 4.2.1), cross-view action anticipation & planning (Sec. 4.2.2) and cross-view reference skill assessment (Sec.4.2.3) in this section. Other benchmarks can be found in the supplementary material.

using the EgoVLP [72] 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 [60, 72, 122, 130, 138, 158] are generally pre-trained on either exocentric [3, 82] or egocentric [30] datasets, only revealing weak cross-view bridging ability. Another line of works explores cross-view learning [2, 33, 135, 141] by either transferring the knowledge from one view to the other [69] or training view-invariant video understanding models [124, 143]. 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 [72] and LaViLa [158], or exocentric videos, *i.e.,* InternVideo [129]. 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 [72]	✗	28.8	27.2	32.1	28.9
LaViLa [158]	✗	22.6	24.9	28.7	25.7
InternVideo [129]	✗	27.0	21.2	30.6	21.7
EgoVLP [72]	✓	28.8	29.7	31.5	28.9
LaViLa [158]	✓	21.9	21.4	30.3	25.9
InternVideo [129]	✓	**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 [18]. 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 [30].

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 [16, 52, 84, 104, 111, 131, 145]. We adopt CLIP [97] + TA3N [14] as the baseline model. The second direction entails knowledge distillation (KD) [28, 32, 87], 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 [69] 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 [25, 54, 106, 117] on multi-view action understanding mainly focus on synchronized multi-view videos. Some work studied transferring knowledge [69] from one view to the other or training the view-invariant [124, 143, 159] 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 [22, 58] 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 [23]. 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 [114], 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, 71, 73, 91, 115, 149]. We adopt a more general approach of pairwise ranking since no absolute score is available in most real-world skills [7, 70]. Previous works in this direction only use a pair of videos that are in either egocentric view [22] or in exo view [23, 70]. 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 [22, 23] 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 [76, 133]. 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 [22]	✗	74.79	75.98	77.54	76.85
RAAN [23]	✗	78.45	78.52	82.53	79.09
Who's better [22]	✓	77.91	76.70	79.13	77.02
RAAN [23]	✓	82.08	79.37	83.92	79.36
Ego pairs + Exo					
RAAN [23] + RN	✗	77.92	77.09	81.54	78.26
RAAN [23] + TL	✗	78.81	79.46	82.50	78.73
RAAN [23] + RN	✓	82.14	79.32	83.51	**79.44**
RAAN [23] + 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.

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

[3] 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

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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

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

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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1, 2, 4, 5, 7

[19] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 2

[20] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009. 2

[21] 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

[22] 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

[23] 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

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

[25] 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

[26] 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

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

[28] 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

[29] 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

[30] 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

[31] 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

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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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.

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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

[63] 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

[64] 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

[65] 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

[66] 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.

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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

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

[76] 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

[77] 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

[78] 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

[79] 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

[80] 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

[81] 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

[82] 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

[83] 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

[84] 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

[85] 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

[86] 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

[87] 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

[88] 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

[89] 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

[90] 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

[91] 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

[92] 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

[93] 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

[94] 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

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

[96] 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

[97] 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

[98] 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

[99] 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

[100] 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

[101] 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

[102] 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

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

[104] 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

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

[106] 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

[107] 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

[108] 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

[109] 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

[110] 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

[111] 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

[112] 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

[113] 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

[114] 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

[115] 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

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

[117] 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

[118] 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

[119] 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

[120] 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

[121] 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

[122] 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

[123] 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

[124] 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

[125] 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

[126] 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

[127] 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

[128] 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

[129] 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

[130] 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

[131] 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

[132] 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

[133] 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

[134] 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

[135] 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

[136] 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

[137] 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

[138] 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

[139] 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

[140] 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

[141] 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

[142] 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

[143] 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

[144] 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

[145] 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

[146] 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

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

[148] 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

[149] 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

[150] 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

[151] 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

[152] 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

[153] 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

[154] 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

[155] 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

[156] 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

[157] 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

[158] 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

[159] 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

[160] 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