EgoExo-Fitness: Towards Egocentric and Exocentric Full-Body Action Understanding

Yuan-Ming Li[1,3,‡†], Wei-Jin Huang[1,3,†], An-Lan Wang[1,3,†], Ling-An Zeng[3,4], Jing-Ke Meng[1,3,*], and Wei-Shi Zheng[1,2,3,*]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China; [2] Peng Cheng Laboratory, Shenzhen, China; [3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China; [4] School of Artificial Intelligence, Sun Yat-sen University, China

Abstract. We present EgoExo-Fitness, a new full-body action understanding dataset, featuring fitness sequence videos recorded from synchronized egocentric and fixed exocentric (third-person) cameras. Compared with existing full-body action understanding datasets, EgoExo-Fitness not only contains videos from first-person perspectives, but also provides rich annotations. Specifically, two-level temporal boundaries are provided to localize single action videos along with sub-steps of each action. More importantly, EgoExo-Fitness introduces innovative annotations for interpretable action judgement–including technical keypoint verification, natural language comments on action execution, and action quality scores. Combining all of these, EgoExo-Fitness provides new resources to study egocentric and exocentric full-body action understanding across dimensions of "what", "when", and "how well". To facilitate research on egocentric and exocentric full-body action understanding, we construct benchmarks on a suite of tasks (*i.e.*, action classification, action localization, cross-view sequence verification, cross-view skill determination, and a newly proposed task of guidance-based execution verification), together with detailed analysis. Data and code are available at `https://github.com/iSEE-Laboratory/EgoExo-Fitness/tree/main`.

Keywords: Egocentric video dataset · Full-body action understanding · Fitness practising · Interpretable action judgement

1 Introduction

Imagine that one day you put on your smart eyewear and perform fitness activities. Virtual coach embedded in the eyewear can provide feedback on <u>*what*</u>*,* <u>*when*</u>*, and* <u>*how well*</u> *you performed the action.* Such a vision draws a scenario in the next generation of AI-assisted fitness exercise, which requires the AI agent to have the ability of egocentric full-body action understanding (EgoFBAU).

However, existing full-body action datasets [6, 49, 50, 59, 61, 69, 70, 73] are predominantly collected from exocentric (third-person) cameras. The dependency of

‡: Project lead. †: Equal key contributions. ∗: Corresponding authors.
Emails: {liym266, wanganlan}@mail2.sysu.edu.cn; mengjke@gmail.com; wszheng@ieee.org.

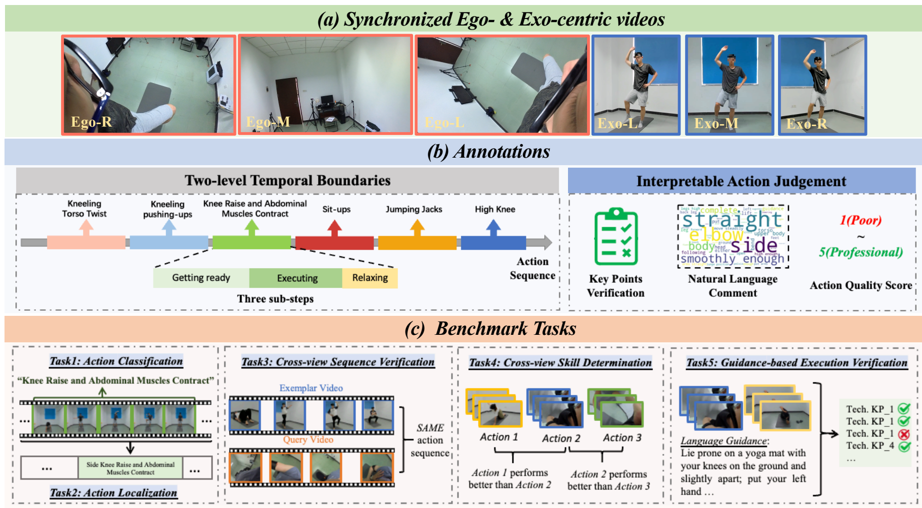

Fig. 1: An Overview of our work. (a) We introduce a new video dataset, namely EgoExo-Fitness, which features synchronized egocentric and exocentric videos of fitness activities to support future work on egocentric full-body action understanding. (b) EgoExo-Fitness provides abundant annotations, including two-level temporal boundaries and interpretable action judgement. (c) We benchmark EgoExo-Fitness on five relevant tasks. Zoom in for the best view.

fixed exocentric cameras limits the technical practicality in a more flexible manner. For instance, it is much more convenient to put on an embodied recording device than to spend time locating a fixed camera. Inspired by the emerged community of egocentric vision [51], we ask, can we embed the virtual coach on your smart eyewear? More generally, **how can we achieve egocentric full-body action understanding?**

By looking at the field of egocentric video understanding, we find that egocentric full-body action understanding is yet to be well explored due to the lack of datasets. Existing egocentric video datasets primarily focus on interactive actions like desktop works [2,10–12,22,28,36,57,67] (*e.g.*, cooking and assembling) and daily interaction [13,23,46,52,60,68] (*i.e.*, interacting with daily objects or humans). The other branch of egocentric datasets [25,33,71,77] mainly focuses on body pose estimation and reconstruction rather than understanding full-body action from other dimensions (*e.g.*, verifying the consistency of action sequences and assessing the execution of action).

To pave the road for future research on full-body action understanding, we focus on fitness activities and present EgoExo-Fitness, a new multi-view full-body action understanding dataset. An overview of EgoExo-Fitness is shown in Fig. 1(a&b). The characteristics of EgoExo-Fitness are as follows:

– Firstly, on data collection, EgoExo-Fitness features a diverse range of fitness sequence videos recorded by synchronized egocentric and exocentric cameras with various directions;

- Secondly, it provides two-level temporal boundaries to localize a single fitness action as well as sub-steps of each action.
- More importantly, EgoExo-Fitness introduces annotations on interpretable action judgement, including technical keypoint verification, natural language comment, and quality score for each single action execution. To our knowledge, no previous dataset contains such annotations on action judgement.

Combining all of these, EgoExo-Fitness spans 32 hours with 1276 cross-view action sequence videos featuring more than 6000 single fitness actions. With synchronized ego-exo videos and rich annotations, EgoExo-Fitness provides new resources to study egocentric and exocentric full-body action understanding across dimensions of "what", "when", and "how well".

To facilitate research on the line of ego- and exo-centric full-body action understanding, as shown in Fig. 1(c), we conduct benchmarks on a suite of tasks, including: *Action Classification*, *Action Localization*, *Cross-View Sequence Verification*, and *Cross-View Skill Determination*. More importantly, to further address interpretable action guiding and action assessment, we propose *Guidance-based Execution Verification*, which aims to infer whether the execution of an action satisfies technical keypoints. Extensive experiments not only evaluate the effectiveness of baseline methods on the benchmark tasks but also pose several challenges for future research.

In summary, the contributions of our work are as follows: 1) We present EgoExo-Fitness, a new full-body action understanding dataset featuring fitness sequence videos recorded from synchronized ego- and exo-centric cameras; 2) We introduce rich annotations on EgoExo-Fitness, including two-level temporal boundaries and novel annotations of interpretable action judgement; 3) We construct benchmarks on five relevant vision tasks, including the newly introduced Guidance-based Execution Verification and extensive experimental analysis.

We expect our dataset and findings can inspire future work on egocentric and exocentric full-body action understanding.

2 Related Works

2.1 Revisiting Current Datasets

We will first revisit today's available full-body action understanding datasets and egocentric video datasets. After that, we will introduce the differences between EgoExo-Fitness and today's datasets.

Full-Body Action Understanding datasets. Human body movements contain complex motion patterns and technical skills, presenting a series of challenges for Full-Body Action Understanding (FBAU). To address these challenges, datasets like NTU-RGB+D [44, 58], Human3.6M [29], Diving48 [40] and FineGym [59] are proposed to enable research on recognizing coarse-and-fine human full-body actions. Beyond recognition, datasets like Diving48-SV [53] and RepCount [27] are present to address tasks (*e.g.*, Sequence Verification and Repetitive Action Counting) that require stronger temporal modeling ability. Note that technical full-body action videos (*e.g.*, diving and vaulting) will reflect human skills. Hence, in recent years, datasets for Action Assessment, like AQA-7 [49], FineDiving [70], LOGO [76], are introduced to study the subtle skill differences between action videos. Another branch of datasets [6, 19, 61] focuses on estimating or reconstructing 3D human poses from full-body action videos, achieving the development of Virtual Reality. Though great progress has been achieved, today's full-body action understanding datasets mainly assume that human full-body action videos are captured by exocentric cameras. Such an assumption limits further exploration in more flexible settings. Moreover, some datasets (*e.g.*, WEAR [5] and 1st-basketball [3]) propose to understand sports and fitness activities from egocentric viewpoints. However, these datasets are limited by their scales and task-specific annotations.

Egocentric Video Understanding Datasets. Egocentric Video Understanding has great application values for AR/VR and Robotics. Most existing datasets focus on interactive actions: 1) tabletop activities in kitchen [10–12, 39, 63] or on a static working platform [2, 22, 36, 55, 57]; 2) daily activities interacting with daily objects [28, 31, 37, 46, 60, 68] or individuals [15, 47, 56]. Although recently proposed Ego4D [23] expands beyond interactive activities to a wider variety of daily activities, works on this branch of datasets rarely focus on egocentric full-body action understanding. Another branch of work aims to estimate full-body pose from egocentric videos, and several datasets [1, 33, 38, 66, 77] are released.

Different from existing datasets, EgoExo-Fitness features synchronized egocentric and exocentric videos of full-body fitness actions and provides rich annotations (especially novel annotations of interpretable action judgement) for future research on understanding ego- and exo-centric full-body actions across the dimensions of "what", "when", and "how well".

It is worth noting that a concurrent large dataset, Ego-Exo4D [24], also contains ego-exo full-body (physical) action videos and annotations on *how well an action is performed*. EgoExo-Fitness still has its values: (1) It focuses on a novel scenario (*i.e.*, natural fitness practising); (2) We provides novel annotations (e.g., technical keypoint verification), enabling studies on new tasks (*e.g.*, Guidance-based Execution Verification). We will provide detailed comparisons between Ego-Exo4D and our work in Sec. 3.5 and Appendix A4.2.

2.2 Revisiting Relevant Tasks

In this part, we will present the relationships between the benchmarks of EgoExo-Fitness and relevant tasks. We will further introduce the motivations and set-ups in detail when introducing the benchmarks.

Action Classification & Localization. As the fundamental tasks in video action understanding, action classification [4, 9, 35, 43] and temporal action localization [18, 45, 65, 75] are widely explored in previous work. In our work, we benchmark EgoExo-Fitness on action classification and localization to present the gap and characteristics across views in full-body action understanding.

Sequence Verification. Sequence Verification (SV) [14, 26, 53] is proposed to study the action order of sequential videos under a scenario that precise temporal annotations are not provided. Today's benchmarks on SV rather focus on

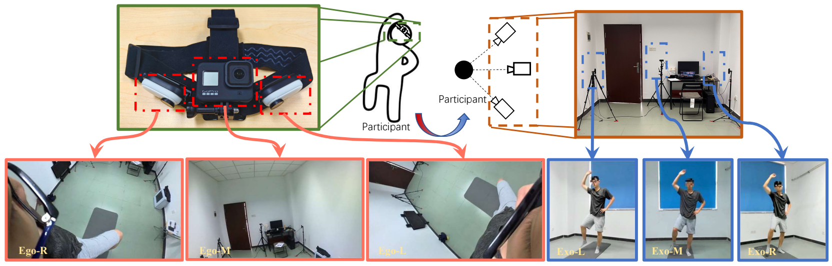

Fig. 2: The setup of our recording system. We capture forward and downward egocentric videos by developing a headset containing three action cameras. To record exocentric videos, three cameras are located at the front, left-front and right-front sides of the actor. Zoom in for the best view.

exocentric-SV (*i.e.*, COIN-SV and Diving48-SV) or egocentric-SV (*i.e.*, CSV). In this work, we present the first benchmark on cross-view sequence verification and provides extensive experimental analysis.

Action Assessment. Existing datasets on Action Assessment (or Skill Determination) are mainly based on videos from either ego-cameras [16, 21] or exo-cameras [17, 20, 49, 50, 69, 70, 73, 74, 76], which leads to single-view assessment ability. Also, today's popular datasets only provide the annotations on action scores or paired rankings, which is unable existing work to directly explore the interpretability of the predicted results. To address the first issue, we introduce the first benchmark on Cross-View Skill Determination. For the second issue, we propose a novel task, Guidance-based Execution Verification, which aims to verify whether the execution of an action satisfies the given technical keypoints.

3 EgoExo-Fitness Dataset

In this section, we introduce the EgoExo-Fitness dataset in detail. We will start by describing the recording system in Sec. 3.1, data collection in Sec. 3.2, and annotations in Sec. 3.3. After that, we present the statistics and comparisons with related datasets in Sec. 3.4 and Sec. 3.5, respectively.

3.1 Recording System

We build a recording system for EgoExo-Fitness to capture action videos from egocentric and exocentric views. Fig. 2 shows the setups of our recording system. For egocentric video capturing, we design a headset with multiple cameras to record videos from forward and downwards views. Specifically, we use a GoPro to record the forward (*i.e.*, Ego-M) view of participants and apply two Insta-Go3 cameras to record the left-downward (*i.e.* Ego-L) and right-downward (*i.e.*, Ego-R) views. For the exocentric cameras, we locate them at the participants' front (*i.e.*, Exo-M), left-front (*i.e.*, Exo-L), and right-front (*i.e.*, Exo-R) sides and ensure they can record full-body actions completely. All cameras are synchronized manually using a timed event that is visible from them.

Table 1: Recorded fitness actions. Abbr.: the abbreviation of the fitness action.

Action types	Abbr.	Action types	Abbr.	Action types	Abbr.
1: Kneeling Push-ups	KPU	5: Shoulder Bridge	SB	10: Jumping Jacks	JJ
2: Push-ups	PU	6: Sit-ups	SU	11: High Knee	HK
3: Kneeling Torso Twist	KTT	7: Leg Reverse Lunge	LRL	12: Clap Jacks	CJ
4: Knee Raise and Abdominal Muscles Contract	KRAMC	8: Leg Lunge with Knee Lift	LLKL		
		9: Sumo Squat	SS		

3.2 Action Sequence & Recording Protocols

Following FLAG3D [61] and HuMMan [6], we select 12 types of fitness actions based on various driving muscles (*i.e.*, *chest*, *abdomen*, *waist*, *hip*, and *whole body*). All selected actions are listed in Tab. 1. In the remaining part of the paper, for convenience, we will use the abbreviation to present the actions. Furthermore, to enrich the temporal diversity of the recorded videos, we define 86 action sequences by randomly combining 3 to 6 different actions. For example, "starting with *Push-ups*, then *Sit-ups*, finally *High Knee*" is an action sequence with three fitness actions. For details of the action sequences, please refer to Appendix A2.

Recording Protocols. Before recording, action sequences will be randomly allocated to the participants. Since we are interested in capturing the natural actions of the participants, we only provide the text guidance in advance. During recording, the participants are asked to put on the headset and continuously complete all actions in the allocated action sequence. For each action, the participants are required to repeat it at least 4 times.

3.3 Annotations

To support future work on EgoFBAU, EgoExo-Fitness provides annotations for two-level temporal boundaries and interpretable action judgement.

Two-level Temporal Boundaries. To enable studies on action boundaries and action orders, we adopt a two-stage strategy to collect the annotations for the two-level temporal structures of each instance. To begin with, given an action sequence video (containing 3 to 6 continuous actions) from any camera view, annotators are asked to accurately locate the start and end time (*e.g.*, t_{st} and t_{ed} in Fig. 3(a)) of each complete action so that a single action video can be obtained. After that, for each single action video, the annotators are asked to separate the video into three steps(*i.e.*, *Getting ready*, *Executing* and *Relaxing*) and annotate the start and end time (*e.g.*, t'_{st} and t'_{ed} in Fig. 3(a)) of the *Executing* steps.

Interpretable Action Judgement. Our motivation for providing this series of annotations is two-fold. First, it is easy for human experts to compare an action video and the text guidance to conclude whether the actor followed the guidance or not and point out which technical keypoint in the text guidance is missed during the execution. However, such ability has rarely been studied in existing video-language answering and video-language retrieval works. Second, although Action Assessment [16, 17, 41, 49, 50, 69, 70] has been studied for many years, and great progress has been achieved, interpretable action assessment

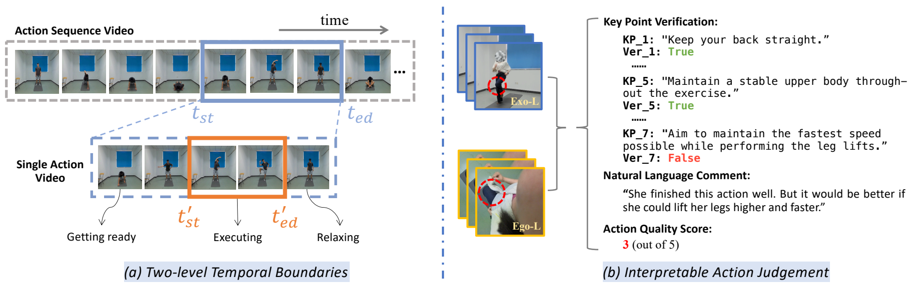

Fig. 3: Overview of annotations setups. (a) Two-level temporal boundaries are provided. Specifically, 1^{st}-level boundaries (t_{st} and t_{ed}) localize the single actions from the action sequence video (obtaining single action videos). After that, 2^{nd}-level boundaries ($t^{'}_{st}$ and $t^{'}_{ed}$) separate every single action into three sub-steps (i.e., getting ready, executing, and relaxing). **(b)** EgoExo-Fitness contains three types of annotations on action judgement, including keypoint verification (KP: keypoint; Ver: verification result), natural language comments, and action quality scores. Zoom in for the best view.

based on language annotations has never been explored due to the lack of well-collected dataset. Also, existing action assessment work is limited to ego-only or exo-only scenarios due to the collecting manner of the datasets.

To address these issues, we develop a web-based annotation tool for EgoExo-Fitness, and collect three categories of interpretable action judgement annotations (*i.e.*, *Technical keypoints Verification*, *Natural Language Comment*, and *Action Quality Score*) step by step. We will introduce the details as follows.

(1) ***Technical Keypoint Verification.*** A paragraph of text guidance on fitness action can be divided into several technical keypoints. By following the keypoints, one can achieve the goal of exercise while avoiding physical injury. In EgoExo-Fitness, we provide the verification annotations on the keypoints for located single actions in the following three steps. First, following FLAG3D [61], we provide a paragraph of text guidance for recorded actions. Second, we prompt LLM (*e.g.*, ChatGPT) to separate the guidance into several keypoints. Third, we ask the annotators to verify an action by comparing the execution with the technical keypoints. Given an action and a technical keypoint, if the action satisfies the keypoint, an annotation of "True" will be noted (otherwise "False").

(2) ***Natural Language Comment.*** After verifying the technical keypoints, the annotators are asked to write a paragraph of natural language comment on how well the participant finished the action. We require that the comments should reflect the verification results from the previous step. Additionally, annotators are asked to write a few sentences on how to improve the movements following their subjective appraisals.

(3) ***Action Quality Score.*** Finally, the annotators are asked to score the actions from 1 to 5 (worst to best) based on the technical keypoint verifications and comments they have made.

Fig. 3(b) gives an example of the annotations on interpretable action judgement. As shown in the frames cropped from ego- and exo-centric videos, the

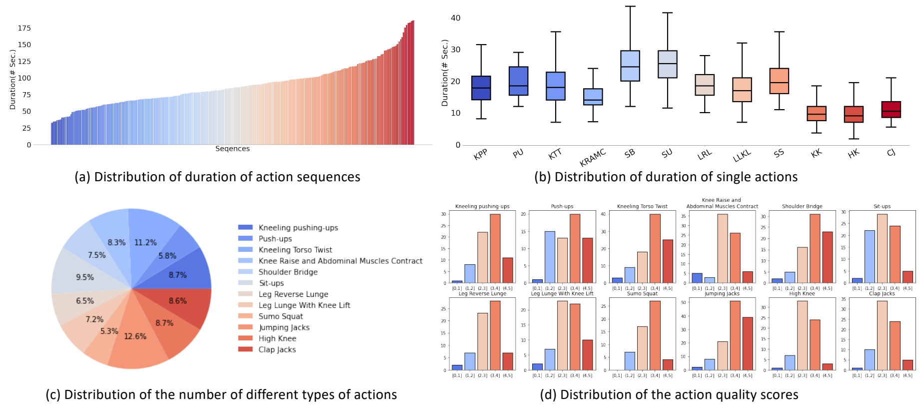

Fig. 4: Statistics of the proposed EgoExo-Fitness dataset.

participant is executing "high knee". Though generally performed well, it can be observed from the video that her legs are not lifted high and fast enough (*i.e.*, *red circles* on the cropped frames). Therefore, relevant keypoints will be verified and annotated as *False* (*i.e.*, KP_7 in Fig. 3(b)). Besides, natural language comments on the execution will be provided together with improvement advice (*i.e.*, "It would be better if she could lift her legs higher and faster"). Finally, a subjective action quality score (*i.e.*, 3) is annotated by the annotator.

To ensure the annotation quality, for each single action video, we employ at least two human experts to provide interpretable action judgement annotations.

3.4 Statistics

Number of recordings and Duration. EgoExo-Fitness collects 1276 cross-view action sequence videos from 86 action sequences, which spans about 32 hours. With two-level temporal boundaries, 6131 single actions are located, Fig. 4(a & b) present the duration distribution of action sequence videos and single action videos. The duration of action sequence videos is widely distributed between 33 and 186 seconds, and most actions last from 10 to 30 seconds. The distribution of the number of different types of action is shown in Fig. 4(c), where "Jumping jacks" takes up the highest proportion of takes (*i.e.*, 12.6%), and the action takes up the fewest proportion of takes is "Sumo Squat" (*i.e.*, 5.3%).

Action Quality Score Distribution. We also analyze the distribution of action quality scores for each type of action in Fig. 4(d). Here, the score for each single action is calculated by averaging the scores annotated by all annotators.

3.5 Comparison with Related Datasets

We compare EgoExo-Fitness with popular ego- and exo-centric full body action understanding datasets Tab. 2. EgoExo-Fitness is the first dataset that features synchronized exo- and ego-centric videos to address egocentric full-body action understanding across dimensions of "what", "when" and "how well". Additionally,

Table 2: Comparison with related datasets. We compare existing datasets on scenarios, annotations and durations. For fair comparison, we select a subset of Ego4D with scenarios of technical full-body action (*e.g.*, dancing, climbing, working-out). SI: Social Interaction. ⋇: MTL-AQA and Ego4D contains captions and narrations on *what have happened* rather than *how well an action has been done* as in EgoExo-Fitness.

Datasets	Public.	Scenarios	Ego	Exo	Step	Text guidance	Keypoint verification	Comment	Score	Duration
Exocentric full-body action datasets										
MTL-AQA [50]	CVPR20	Diving		✓				⋇	✓	1.6h
FineGym [59]	CVPR20	Sports		✓	✓					161h
FineDiving [70]	CVPR22	Diving		✓	✓				✓	3h
FLAG3D(virtual) [61]	CVPR23	Fitness		✓		✓				∼185h
FLAG3D(real) [61]	CVPR23	Fitness		✓		✓				69h
Egocentric full-body action datasets										
1st-basketball [3]	ICCV17	Basketball	✓						✓	10.3h
Ego4D(tfba) [23]	CVPR22	Daily	✓		✓			⋇		172h
WEAR [5]	ArXiv23	Fitness	✓		✓					15h
EgoExo-Fitness(Ours)		Fitness	✓	✓	✓	✓	✓	✓	✓	32h

Table 3: Comparison with the concurrent Ego-Exo4D [24] dataset. The proposed EgoExo-Fitness collects videos of a new scenario and augments data with novel annotations. For fair comparison, scenarios of full-body actions are considered.

Datasets	Scenarios	Step	Text guidance	Keypoint verification	Comment	Score	Duration
Ego-Exo4D v2 [24]	Basketball				✓	✓	74h
	Climbing				✓	✓	93h
	Soccer				✓	✓	66h
	Dancing				✓	✓	106h
EgoExo-Fitness(Ours)	Fitness	✓	✓	✓	✓	✓	32h

EgoExo-Fitness introduces novel annotations on interpretable action judgement (*i.e.*, keypoint verifications and comments on how well an action is performed), which make EgoExo-Fiteness different from existing datasets. With synchronized videos and rich annotations, EgoExo-Fitness provides new resources for studying view characteristics, cross-view modeling, and action guiding. It is also notable that for a fair comparison, we select a subset from Ego4D [23], which includes scenarios of technical full-body actions (e.g., dancing, and working-out).

Note that a recent proposed large-scale dataset Ego-Exo4D [24] also contains *full-body (physical)* action videos collected by synchronized ego-exo cameras. Besides, they both attend to *how well* an action is performed and propose novel corresponding annotations. What makes our dataset different from Ego-Exo4D lines in the following ways (also shown in Tab. 3): **(1)** ***New scenario is focused on***. We focus on the scenario of natural fitness practising and collect dynamic action sequence videos (containing 3 to 6 different actions). However, in Ego-Exo4D, a video is only associated with one type of action/task. This makes our dataset better suited for ego-exo full-body action studies on action boundaries and orders. **(2)** ***Novel annotations are provided***. We provide text guidance and technical keypoints verification (both NOT included in Ego-Exo4D), which offer another branch of intuitive and detailed identifications of *what is done well* and *what can be improved* in executions than expert commentary in Ego-Exo4D. Such annotations enable the pioneering exploration of interpretable action assessment. **(3)** ***Other unique characteristics***. We also

Table 4: Action classification benchmark results on different models. Top-1 accuracies are reported for different models with different pretraining strategies. **Bolded** and underlined values indicates the best and 2-nd best results, respectively.

Train on	Models	Pretrain	Test on Exo	Test on Ego	Train on	Models	Pretrain	Test on Exo	Test on Ego
Exo	I3D [9]	K400 [32]	0.9194	0.0927	Exo	EgoVLP [42]	Ego4D [23]	0.8940	0.0893
Ego			0.1025	0.7469	Ego			0.0887	0.7977
Ego & Exo			0.8963	0.7266	Ego & Exo			0.8986	0.7932
Exo	TSF [4]	K600 [8]	**0.9274**	0.0836	Exo	TSF [4]	K600 [8] + Ego-Exo4D [24]	0.8825	0.0814
Ego			0.1417	0.7932	Ego			0.1601	**0.8000**
Ego & Exo			0.8894	0.7842	Ego & Exo			0.8975	0.7840

provide videos captured from two downward ego-cameras for capturing more body details in movements, and annotations of two-level temporal boundaries to enable benchmark constructions.

For more details about dataset comparisons, please refer to Appendix A4.

4 Benchmarks

With synchronized ego-exo videos and rich annotations, EgoExo-Fitness can provide resources for studies of view characteristics, cross-view modeling, and action guiding. To benefit future research of these directions on EgoExo-Fitness, we conduct benchmarks on *Action Classification* (Sec. 4.1), *Cross-View Sequence Verification* (Sec. 4.2), and a newly proposed *Guidance-based Execution Verification* (Sec. 4.3). EgoExo-Fitness also supports *Action Localization* and *Cross-View Skill Determination*, which are presented in Appendix A3.

4.1 Action Classification

We select Action Classification [35], the fundamental task of video action understanding, to study view gap and view characteristics on EgoExo-Fitness.

Task Setups. We share the same task setups with previous works on action classification, *i.e.*, to predict the type of fitness action given a trimmed single action video from either ego-or-exo viewpoint.

Baseline Models. We apply three baseline models in Action Classification benchmark: (1) I3D [9] pre-trained on K400 dataset [32]; (2) TimeSformer(TSF) [4] pre-trained on K600 [8] and Ego-Exo4D(EE4D) [24] datasets; (3) EgoVLP [42] pre-trained on Ego4D(E4D) [23] dataset.

Experiment Results. Top-1 accuracies of different models are reported in Tab. 4. We analyze the results in the following aspects. **(1)** ***Impacts of pre-training***. Among all results, TSF and I3D pre-trained on Kinetics datasets achieve the best two performances (0.9274 and 0.9194) on exocentric videos. Similarly, TSF pre-trained on EE4D performs best (0.8000) on egocentric videos, closely followed by the one pre-trained on E4D (0.7977). Such results are attributed to view-related pre-training datasets (*i.e.*, Kinetics are exocentric datasets; E4D and EE4D consist of various egocentric videos). **(2)** ***Analysis on the***

view gap. Not surprisingly, models trained ego-only or exo-only data suffer from a significant performance drop on cross-view testing. Additionally, we find that mixing up cross-view data (Ego & Exo) for training does not always bring performance improvement. For I3D and TSF pre-trained on Kinetics datasets, performance drops on both egocentric and exocentric data. For TSF pre-trained from E4D and EE4D, only performance on exocentric data obtains improvement when mixing up cross-view data for training. Such results indicate a great domain gap between ego-videos and exo-videos. **(3)** ***Why do models perform worse on ego-videos?*** From Tab. 4, we also observe that models always perform worse on ego-videos than on exo-videos. We think this is because it is easier to observe similar action patterns from egocentric videos, which confuse models. Another reason is that it is more difficult to find discriminating clues from the Ego-M camera. Appendix A3.1 provides more analysis supporting these views.

4.2 Cross-view Sequence Verification

Sequence Verification (SV) [14, 26, 53] is proposed to verify the action order consistency of sequential videos under a scenario where precise temporal annotations are not provided, which shows great potential in video abstraction, industrial safety, and skill studying. Existing SV datasets [53] are collected either from exocentric or egocentric cameras, which constrains existing studies in an inner-view manner. However, it is desirable to study whether a model can perform promising verification of two videos from egocentric and exocentric views. For instance, during our daily fitness exercises, an AI assistant in our eyewear can remind us whether we have missed any exercise program by verifying the sequence of exocentric expertise exemplar videos and the self-recorded egocentric videos. Hence, we extend the traditional SV to Cross-View SV (CVSV).

Task Setups. CVSV aims to verify whether two fitness sequence videos have identical procedures. Two action sequence videos executing the same steps in the same order form a positive pair; otherwise, they are negative. The method should give a verification distance between each video pair based on the video representations, and give the prediction by thresholding the distance.

CVSV is more challenging than traditional SV because videos can be shot from either egocentric or exocentric cameras. Hence, it is crucial for models to learn retrievable (or translatable) representations across views. More formal task setups will be introduced in Appendix A3.2.

Baseline model. We use the state-of-the-art SV model CAT [53] to conduct experiments. More details about the CAT will be introduced in Appendix A3.2.

Metrics. (1) ***AUC***: Following existing works [14, 26, 53], we first adopt the Area Under ROC Curve to evaluate the performance. **(2)** ***Rank 1 & mAP***: To further study the relations among learned representations, we borrow the idea of image retrieval [7, 30, 34] and use Rank-1 and mAP to evaluate CVSV models.

Experiment Results. Benchmark results are reported in Tab. 5 and Tab. 6. We analyze the results in the following aspects. **(1)** ***Influence of cross-view training data.*** As the first attempt, we wonder how cross-view training data

Table 5: Cross-view sequence verification results on various training sources. We find that mixing-up all training pairs will benifit exo-only and ego-exo SV, but bring declines on ego-only setting. "View1-View2" indicates that the model takes video pairs with one video from *View1* and the other from *View2*. "View1→View2" indicates taking videos from *View1* to retrieve videos from *View2*.

Train on			AUC			Rank 1				mAP			
Exo-Exo	Ego-Ego	Exo-Ego	Ego-Ego	Exo-Exo	Ego-Exo	Ego→Ego	Exo→Exo	Ego→Exo	Exo→Ego	Ego→Ego	Exo→Exo	Ego→Exo	Exo→Ego
✓			0.532	0.800	0.577	0.165	0.583	0.094	0.092	0.087	0.374	0.117	0.080
	✓		**0.803**	0.487	0.480	**0.620**	0.071	0.040	0.021	**0.325**	0.064	0.057	0.065
		✓	0.761	0.813	0.744	0.539	0.646	0.296	0.363	0.275	0.394	0.228	0.237
✓	✓		0.751	0.814	0.743	0.556	0.629	0.286	**0.383**	0.275	0.402	0.223	0.238
✓	✓	✓	0.759	**0.822**	**0.776**	0.572	**0.663**	**0.300**	0.367	0.281	**0.406**	**0.247**	**0.247**

Table 6: Cross-view sequence verification results on imbalanced training data. We gradually prune egocentric videos from training data and we find that it is challenging to perform cross-view sequence verification with imbalanced data.

Prune Rate	AUC	Rank 1		mAP		Prune Rate	AUC	Rank 1		mAP	
	Exo-Ego	Ego→Exo	Exo→Ego	Ego→Exo	Exo→Ego		Exo-Ego	Ego→Exo	Exo→Ego	Ego→Exo	Exo→Ego
100%	0.5768	0.0943	0.0917	0.1174	0.0798	30%	0.7072	0.1751	0.2917	0.1824	0.1905
70%	0.6562	0.1077	0.0750	0.1412	0.1109	0%	**0.7755**	**0.2997**	**0.3667**	**0.2470**	**0.2468**

will influence the performance. Hence, we separate all training video pairs into three parts (*i.e.*, *Exo-Exo*, *Ego-Ego* and *Exo-Ego*) to study how cross-view training data would influence the performance. The results are shown in Tab. 5. First, we observe that combining all training pairs will benefit performance on exo-only and ego-exo pairs but bring a performance drop on ego-only pairs, which further indicates the domain gap between different views. Furthermore, compared with 0.8033 on ego-only data and 0.8221 on exo-only data, the best SV performance on ego-exo data is 0.7755, which indicates that cross-view sequence verification is a challenging task. Retrieval results also support this conclusion. Cross-view retrieval achieves much poorer performance (0.3 on Rank 1 and 0.25 on mAP) compared with inner-view retrieval. **(2)** ***How many egocentric data is needed for CVSV?*** In practical application, it is much easier to collect exocentric videos than egocentric videos. Hence, it is desirable to study if a CVSV model can achieve superior performance with limited egocentric training videos. To the end, we gradually prune egocentric videos from the training set (*i.e.*, 0%, 30%, 70% and 100%) and evaluate the performance. As shown in Tab. 6, when gradually prone training data of egocentric videos, the performance drops on all metrics, posing a great challenge for future study on settings with unbalanced (*i.e.*, limited egocentric videos and rich exocentric videos) cross-view data.

4.3 Guidance-based Execution Verification

Existing works in Action Assessment mainly focus on predicting the final score of an action video or a pair-wise ranking between a pair of videos. However, in real-world action guiding scenarios, providing interpretable feedback is more valuable than giving a score or a ranking. For example, our fitness coach will tell us which technical keypoints are not satisfied in our executions, which will not only explain how well we have performed but also let us know how to improve. However, such an ability has never been explored in action assessment. To address this issue,

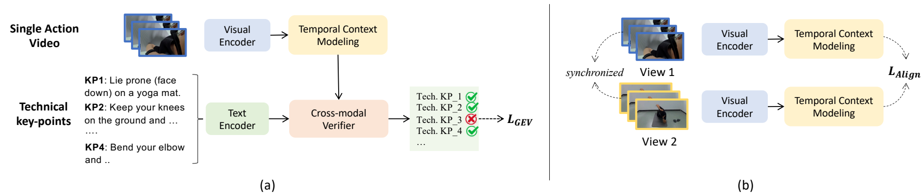

Fig. 5: Overview of GEVFormer. (a) GEVFormer takes an action video and technical keypoints as input, and output the verification results on each keypoint. (b) During training, a synchronized video alignment loss is adopted to force the model to obtain consistent representations across synchronized videos from various views.

we make the first attempt to study interpretable action assessment and propose a novel task termed Guidance-based Execution Verification (GEV).

Task Setups. Given a set of technical keypoints in text as the guidance, the goal of GEV is to verify whether the execution of an action satisfies the keypoints in the guidance. Formally, given an action video v and n technical keypoints $Q = \{q_1, q_2, ..., q_n\}$, a model F is asked to perform an n-way score prediction $P = \{p_1, ..., p_n\} = F(v, Q)$, where p_i represents the verification score of the i-th keypoint. The higher the p_i, the more likely the action satisfies the i-th keypoint. During inference, a threshold τ is adopted to verify whether the action satisfies the keypoints. If $p_i > \tau$, the model predicts the action "satisfies" the i-th keypoint. Otherwise, the model returns a result of being "unsatisfies".

Baseline Model. To better address GEV, we introduce a transformer-based [64] model named GEVFormer, which tasks a single action video and the corresponding technical keypoints as input and outputs the verification results for each keypoint. As shown in Fig. 5(a), video and keypoints are fed into the visual and text encoder to obtain visual and text embeddings. After that, a Temporal Context Modeling (TCM) module is adopted to model temporal information of visual embeddings further, obtaining enhanced visual embeddings. Finally, text embeddings and enhanced visual embeddings are fed into a Cross-Modal Verifier (CMV) to obtain the results.

During training, a loss for GEV, denoted as L_{GEV}, is adopted to require the model to provide accurate verification results. Besides, in our early experimental attempts on GEV, we have the same observations as in other tasks that simply combining training data from egocentric and exocentric views cannot bring stable performance improvement due to the domain gap between different views. To bridge the gap, as shown in Fig. 5(b), inspired by previous works on cross-view learning [62, 72], we propose to utilize an InfoNCE-based [48] alignment loss, denoted as L_{Align}, to force model to obtain consistent representations across synchronized videos from various views. The overall training loss is written as $L = L_{GEV} + \lambda L_{Align}$, where λ is a hyper-parameter.

In our implementation, visual and text encoders are designed as the image and text encoders of pre-trained CLIP [54] and frozen during training. The TCM

Table 7: Guidance-based execution verification results on different baselines. We report performance on ego- and exo-centric data alone with the average F1-score. "ego/exo" indicates ndependent models are trained on ego-only and exo-only data. "ego+exo" indicates the model is trained on both ego-and-exo data.

Methods	Exo			Ego			Avg
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Random	0.3178	0.2329	0.5000	0.3178	0.2329	0.5000	0.3178
Distribution Prior	0.2323	0.2323	0.2323	0.2323	0.2323	0.2323	0.2323
CLIP-GEV(ego+exo)	0.5080	0.5362	0.4657	0.4780	0.5401	0.4094	0.4881
GEVFormer(ego/exo) w/o alignment	**0.5474**	0.5541	0.5408	0.5161	0.5067	0.5259	0.5318
GEVFormer(ego+exo) w/o alignment	0.5282	0.5502	0.5080	<u>0.5248</u>	0.5570	0.4960	<u>0.5265</u>
GEVFormer(ego+exo)	<u>0.5452</u>	0.5219	0.5707	**0.5425**	0.5186	0.5687	**0.5439**

is designed as a Transformer Encoder and the CMV contains a Transformer Decoder together with a linear evaluator. See more details in Appendix A3.3.

Experiment Results. We compare GEVFormer with four other naive methods and invariants: **(1)** ***Random***: Randomly predict an action satisfies a technical keypoint with 50% probability; **(2)** ***Distribution Prior***: Randomly predict an action satisfies a keypoint with the distribution prior; **(3)** ***CLIP-GEV***: Simply concatenate average-pooled visual embedding and text embeddings extracted by CLIP [54] and feed it into a linear evaluator to predict the results; **(4)** ***GEV-Former w/o alignment***: Ablate L_{Align} from GEVFormer. To evaluate the methods, we adopt the *Precision*, *Recall*, and *F1-score*. Here we regard "unsatisfies" as the positive label since samples with "satisfies" labels take up a much higher proportion than those with "unsatisfies" labels in EgoExo-Fitness.

As shown in Tab. 7, GEVFormer outperforms all naive baselines. Besides, compared with the variants of GEVFormer, we have the same findings as on the other tasks that jointly training models on ego- and exo-centric data will not bring stable improvement (achieving 0.0087 improvement on egocxentric data with 0.0192 drop on exocentric data). Surprisingly, when further adopting the L_{Align}, GEVFormer achieves the best performance on egocentric data with 0.5425 F1-score, suffering only a 0.0022 performance drop on exocentric data.

5 Conclusion

We believe that studying egocentric full-body action understanding will benefit the development of AI-assistant. To enable this line of research, we focus on the scenario of fitness exercise and guiding, and introduce EgoExo-Fitness. With a diverse range of synchronized ego- and exo-centric fitness action sequence videos and rich annotations on temporal boundaries and interpretable action judgement, EgoExo-Fitness provides new resources for egocentric and exocentric full-body action understanding. To facilitate future research on EgoExo-Fitness, we construct benchmarks on five relevant tasks. Through experiment analysis, we evaluate the performance of baseline models and point out several interesting problems that await future research (*e.g.*, how to better address cross-view modeling with unbalanced data; how to leverage synchronized exocentric data to achieve better performance).

Acknowledgments

This work was supported partially by the National Key Research and Development Program of China (2023YFA1008503), NSFC(U21A20471, 62206315), Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085, 2024A1515010101), Guangzhou Basic and Applied Basic Research Scheme(2024A04J4067). The authors thank Kun-Yu Lin for the valuable discussions. The authors also thank anonymous reviewers and ACs for their constructive suggestions.

References

1. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: Unrealego: A new dataset for robust egocentric 3d human motion capture. In: Proceedings of the European Conference on Computer Vision. pp. 1–17. Springer (2022)
2. Bansal, S., Arora, C., Jawahar, C.: My view is the best view: Procedure learning from egocentric videos. In: Proceedings of the European Conference on Computer Vision. pp. 657–675. Springer (2022)
3. Bertasius, G., Soo Park, H., Yu, S.X., Shi, J.: Am i a baller? basketball performance assessment from first-person videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2177–2185 (2017)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (2021)
5. Bock, M., Moeller, M., Van Laerhoven, K., Kuehne, H.: Wear: A multimodal dataset for wearable and egocentric video activity recognition. arXiv preprint arXiv:2304.05088 (2023)
6. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., et al.: Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In: Proceedings of the European Conference on Computer Vision. pp. 557–577. Springer (2022)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
8. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision. pp. 720–736 (2018)
11. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision. arXiv preprint arXiv:2006.13256 (2020)
12. DelPreto, J., Liu, C., Luo, Y., Foshey, M., Li, Y., Torralba, A., Matusik, W., Rus, D.: Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. Advances in Neural Information Processing Systems **35**, 13800–13813 (2022)

13. Diete, A., Sztyler, T., Stuckenschmidt, H.: Vision and acceleration modalities: Partners for recognizing complex activities. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops. pp. 101–106. IEEE (2019)
14. Dong, S., Hu, H., Lian, D., Luo, W., Qian, Y., Gao, S.: Weakly supervised video representation learning with unaligned text for sequential videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2447 (2023)
15. Donley, J., Tourbabin, V., Lee, J.S., Broyles, M., Jiang, H., Shen, J., Pantic, M., Ithapu, V.K., Mehra, R.: Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. arXiv preprint arXiv:2107.04174 (2021)
16. Doughty, H., Damen, D., Mayol-Cuevas, W.: Who's better? who's best? pairwise deep ranking for skill determination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6057–6066 (2018)
17. Doughty, H., Mayol-Cuevas, W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7862–7871 (2019)
18. Du, J.R., Feng, J.C., Lin, K.Y., Hong, F.T., Wu, X.M., Qi, Z., Shan, Y., Zheng, W.S.: Weakly-supervised temporal action localization by progressive complementary learning. arXiv preprint arXiv:2206.11011 (2022)
19. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9919–9928 (2021)
20. Gao, J., Zheng, W.S., Pan, J.H., Gao, C., Wang, Y., Zeng, W., Lai, J.: An asymmetric modeling for action assessment. In: Proceedings of the European Conference on Computer Vision. pp. 222–238. Springer (2020)
21. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention Workshop (2014)
22. Ghoddoosian, R., Dwivedi, I., Agarwal, N., Dariush, B.: Weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10128–10138 (2023)
23. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
24. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
25. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4318–4329 (2021)

26. He, T., Liu, H., Li, Y., Ma, X., Zhong, C., Zhang, Y., Lin, W.: Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis. arXiv preprint arXiv:2312.11024 (2023)
27. Hu, H., Dong, S., Zhao, Y., Lian, D., Li, Z., Gao, S.: Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19013–19022 (2022)
28. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., et al.: Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22072–22086 (2024)
29. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2013)
30. Jang, Y.K., Cho, N.I.: Self-supervised product quantization for deep unsupervised image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12085–12094 (2021)
31. Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., Mayol-Cuevas, W.: Epic-tent: An egocentric video dataset for camping tent assembly. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
32. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
33. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: Egohumans: An egocentric 3d multi-human benchmark. arXiv preprint arXiv:2305.16487 (2023)
34. Kim, D., Saito, K., Oh, T.H., Plummer, B.A., Sclaroff, S., Saenko, K.: Cds: Cross-domain self-supervised pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9123–9132 (2021)
35. Kong, Y., Fu, Y.: Human action recognition and prediction: A survey. International Journal of Computer Vision (2022)
36. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10138–10148 (2021)
37. Lei, S.W., Gao, D., Wang, Y., Mao, D., Liang, Z., Ran, L., Shou, M.Z.: Assistsr: Task-oriented video segment retrieval for personal ai assistant. arXiv preprint arXiv:2111.15050 (2021)
38. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)
39. Li, Y., Liu, M., Rehg, J.M.: 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). pp. 619–635 (2018)
40. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 513–528 (2018)
41. Li, Y.M., Zeng, L.A., Meng, J.K., Zheng, W.S.: Continual action assessment via task-consistent score-discriminative feature distribution modeling. IEEE Transactions on Circuits and Systems for Video Technology (2024)

42. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems **35**, 7575–7586 (2022)
43. Lin, K.Y., Ding, H., Zhou, J., Peng, Y.X., Zhao, Z., Loy, C.C., Zheng, W.S.: Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. arXiv preprint arXiv:2403.01560 (2024)
44. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2684–2701 (2019)
45. Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., Bai, X.: End-to-end temporal action detection with transformer. IEEE Transactions on Image Processing **31**, 5427–5441 (2022)
46. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21013–21022 (2022)
47. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9890–9900 (2020)
48. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
49. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 1468–1476. IEEE (2019)
50. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 304–313 (2019)
51. Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G.M., Damen, D., Tommasi, T.: An outlook into the future of egocentric vision. arXiv preprint arXiv:2308.07123 (2023)
52. Possas, R., Caceres, S.P., Ramos, F.: Egocentric activity recognition on a budget. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5967–5976 (2018)
53. Qian, Y., Luo, W., Lian, D., Tang, X., Zhao, P., Gao, S.: Svip: Sequence verification for procedures in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19890–19902 (2022)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
55. Ragusa, F., Furnari, A., Livatino, S., Farinella, G.M.: The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1569–1578 (2021)
56. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14663–14674 (2023)
57. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)

58. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016)
59. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
60. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018)
61. Tang, Y., Liu, J., Liu, A., Yang, B., Dai, W., Rao, Y., Lu, J., Zhou, J., Li, X.: Flag3d: A 3d fitness activity dataset with language instruction. In: CVPR (2023)
62. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Proceedings of the European Conference on Computer Vision. pp. 776–794. Springer (2020)
63. De la Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., Beltran, P.: Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University (2009)
64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
65. Wang, B., Zhao, Y., Yang, L., Long, T., Li, X.: Temporal action localization in the deep learning era: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
66. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023)
67. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20270–20281 (2023)
68. Wong, B., Chen, J., Wu, Y., Lei, S.W., Mao, D., Gao, D., Shou, M.Z.: Assistq: Affordance-centric question-driven task completion for egocentric assistant. In: Proceedings of the European Conference on Computer Vision. pp. 485–501. Springer (2022)
69. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE Transactions on Circuits and Systems for Video Technology **30**(12), 4578–4590 (2019)
70. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2949–2958 (2022)
71. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE Transactions on Visualization and Computer Graphics **25**(5), 2093–2101 (2019)
72. Yu, H., Cai, M., Liu, Y., Lu, F.: What i see is what you see: Joint attention learning for first and third person video co-analysis. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1358–1366 (2019)

73. Zeng, L.A., Hong, F.T., Zheng, W.S., Yu, Q.Z., Zeng, W., Wang, Y.W., Lai, J.H.: Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: Proceedings of the ACM International Conference on Multimedia. pp. 2526–2534 (2020)
74. Zeng, L.A., Zheng, W.S.: Multimodal action quality assessment. IEEE Transactions on Image Processing (2024)
75. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision (2022)
76. Zhang, S., Dai, W., Wang, S., Shen, X., Lu, J., Zhou, J., Tang, Y.: Logo: A long-form video dataset for group action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2405–2414 (2023)
77. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: Proceedings of the European Conference on Computer Vision. pp. 180–200. Springer (2022)