FLAG3D++: A Benchmark for 3D Fitness Activity Comprehension with Language Instruction

Yansong Tang, *Member, IEEE*, Aoyang Liu, Jinpeng Liu, Shiyi Zhang, Wenxun Dai, Jie Zhou, *Senior Member, IEEE*, Xiu Li, *Member, IEEE*, and Jiwen Lu, *Fellow, IEEE*

Abstract—Recent years have witnessed the rapid development of general human action understanding. However, when applied to real-world applications such as sports analysis, most existing datasets are still unsatisfactory, because of the limitations in rich labels on multiple tasks, language instructions, high-quality 3D data, and diverse environments. In this paper, we present FLAG3D++, a large-scale benchmark for 3D fitness activity comprehension, which contains 180K sequences of 60 activity categories with language instruction. FLAG3D++ features the following four aspects: 1) fine-grained annotations of the temporal intervals of actions in the untrimmed long sequences and how well these actions are performed, 2) detailed and professional language instruction to describe how to perform a specific activity, 3) accurate and dense 3D human pose captured from advanced MoCap system to handle the complex activity and large movement, 4) versatile video resources from a high-tech MoCap system, rendering software, and cost-effective smartphones in natural environments. In light of the specified features, we present two new practical applications as language-guided repetition action counting (L-RAC) and language-guided action quality assessment (L-AQA), which aim to take the language descriptions as references to count the repetitive times of an action and assess the quality of action respectively. Furthermore, we propose a Hierarchical Language-Guided Graph Convolutional Network (HL-GCN) model to better fuse the language information and skeleton sequences for L-RAC and L-AQA. To be specific, the HL-GCN performs cross-modal alignments by the early fusion of the linguistic feature and the hierarchical node features of the skeleton-based sequences encoded by the multiple intermediate graph convolutional layers. Extensive experiments show the superiority of our HL-GCN on both L-RAC and L-AQA, as well as the great research value of FLAG3D++ for various challenges, such as dynamic human mesh recovery and cross-domain human action recognition. Our dataset, source code, and trained models are made publicly available at FLAG3D++.

Index Terms—Activity understanding, large-scale benchmark, graph convolutional network, cross-modal fusion

1 Introduction

With the great demand for keeping healthy, reducing high pressure from working, and staying in good shape, fitness activity has become more and more important and popular during the past decades [1], [2], [3], [4], [5]. According to the statistics[1], there are over 200,000 fitness clubs and 170 million club members all over the world. Therefore, it is desirable to advance current intelligent vision systems to assist people effectively in perceiving, understanding, and analyzing various fitness activities in our lives.

In recent years, great achievements have been made for the datasets and techniques on human fitness activity understanding [6], [7], [8], [9], [10], [11], [12], [13], [14], such as classifying action classes like "jumping" and "running". However, when delving into the finer-grained application, *e.g.,* evaluating how well a specific action is performed, the most existing fitness activity datasets [1], [6], [15], [16] might still exhibit limitations in accurately describing fine-grained activities, modeling complex poses, and demonstrating generalizability across diverse scenarios, which make them far from the real-world scenario. In this paper, we introduce FLAG3D++, a comprehensive 3D **F**itness activity dataset complemented by **LAnG**uage instructions. Fig. 1 visually encapsulates our dataset, comprising 180,000 sequences encompassing 60 intricate fitness activities sourced from diverse outlets, including cutting-edge MoCap systems, professional rendering software, and cost-effective smartphones. Notably, FLAG3D++ propels existing datasets forward by addressing four key aspects:

Fine-grained Annotations. We provide fine-grained annotations of the temporal segments of actions in the untrimmed long sequences [17], [18], [19] and how well an action is performed by professional trainers [20], [21], [22], [23], [24], bases on the language instructions. These fine-grained linguistic annotations are critical for in-depth analysis of human movement and the advancement of action understanding systems. Based on these annotations, we further present two new practical applications as language-guided repetition action counting (L-RAC) and language-guided action quality assessment (L-AQA), which specifically aim to take the language instructions as references to count the repetitive times of an action and assess the quality of this action respectively. The L-RAC task is the preliminary stage for the L-AQA task. It divides the action sequence into clear distinct repetitive segments, allowing the action quality assessment task to thoroughly evaluate each segment individually, enabling the model to analyze movements across different repetitions and precisely identify issues with the quality of people's movements.

Detailed Language Instruction. Most existing fitness activity datasets merely provide a single action label or phase for each action [15], [16]. However, understanding fitness activities usually requires more detailed descriptions. We collect a series of sentence-level language instructions for describing each fine-grained movement. Introducing language would also facilitate various research regarding emerging multi-modal applications.

Highly Accurate and Dense 3D Pose. For fitness activi-

• *Y. Tang, A. Liu, J. Liu, S. Zhang, W. Dai, X. Li are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, China 518071. J. Zhou and J. Lu are with the Department of Automation, Tsinghua University, China 100084. (contact e-mail: tang.yansong@sz.tsinghua.edu.cn, lujiwen@tsinghua.edu.cn).*

1. https://policyadvice.net/insurance/insights/fitness-industry-statistics

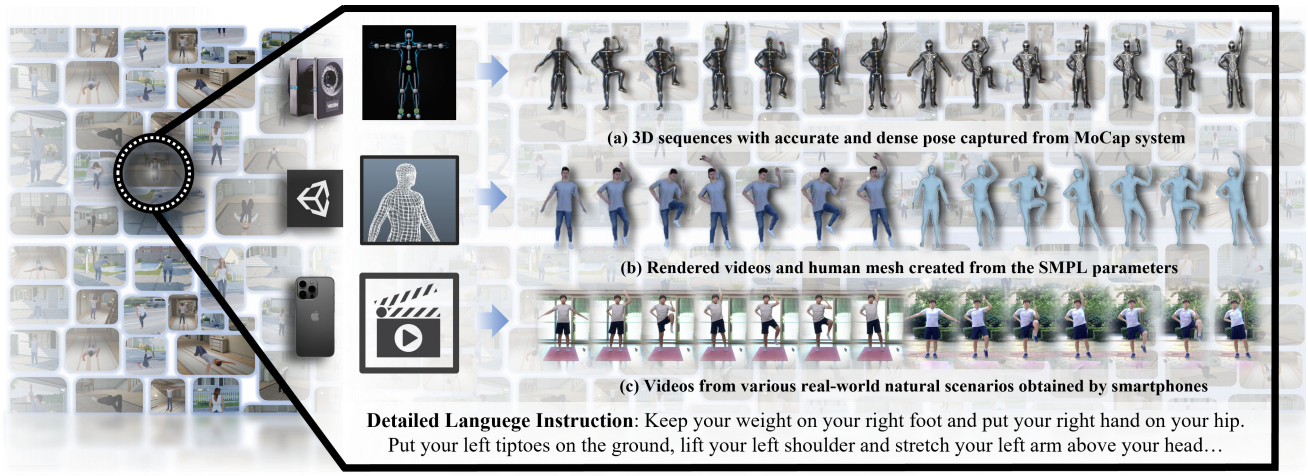


Fig. 1: An overview of the proposed FLAG3D++ dataset, which contains 180K videos for 60 daily fitness activities. Our dataset is comprised of (a) 3D human activity sequences captured from an advanced and precise MoCap system, (b) rendered videos of different people with their SMPL parameters, and (c) real-world videos obtained by cost-effective phones from both indoor and outdoor natural environments. FLAG3D++ also provides a series of detailed and professional sentence-level language instructions for each fitness activity, as well as more fine-grained annotations for action quality assessment and repetition action counting.

ties, there are various poses within lying, crouching, rolling up, jumping, and so on, which involve heavy self-occlusion and large movements. These complex cases bring inevitable obstacles for conventional appearance-based or depth-based sensors to capture the accurate 3D pose, hindering the trained model's capacity to regress the human's pose and shape. To address this, we set up an advanced MoCap system with 24 VICON cameras [25] and professional MoCap clothes with 77 motion markers to capture the detailed and dense 3D pose of the actors during data collection.

Diverse Video Resources. To advance the research directly into a more general field, we collect versatile videos for FLAG3D++. Besides the data captured from the expensive MoCap system [25], [26], [27], we further provide the synthetic sequences with high realism produced by rendering software and the corresponding SMPL [28], [29] parameters. In addition, FLAG3D++ also contains videos from natural real-world environments obtained by cost-effective and user-friendly smartphones.

Moreover, we propose a Hierarchical Language-Guided Graph Convolutional Network (HL-GCN) approach to better integrate language information and guide the model in learning language-related motion representations. Specifically, HL-GCN first utilizes multiple graph convolutional layers [30], [31] to extract hierarchical features of the non-grid skeleton-based activity sequence. Then for the language instructions in FLAG3D++, we observe that they typically describe multiple concepts about the activity, such as the skeleton position, motion velocity, semantic category and so on. These different concepts may not all be best represented by the same level of skeleton-based feature map, *e.g.,* clues for the described skeleton positions may be more evident in lower-level features, and those for a semantic category are more likely to be embedded in high-level features. To this end, rather than directly fusing the linguistic features, we perform the cross-modal alignment in a layer-by-layer manner, which enables the model to maintain awareness of language-guidance information and allows more efficient representation of various linguistic concepts (*e.g.,* position, categories) in different level of the multi-modal features. Extensive experimental results show that our approach achieves state-of-the-art results in both L-RAC and L-AQA.

In addition, to comprehensively grasp other emerging challenges within FLAG3D++, we assess a range of cutting-edge methodologies, establishing benchmarks across diverse tasks encompassing human mesh recovery and skeleton-based action recognition. Our experimental results reveal the following insights which reflect the challenges our FLAG3D++ brings: 1) The existing 3D pose and shape estimation approaches [32], [33], [34], [35] encounter notable limitations, particularly in instances of challenging poses like kneeling and lying, owing to inherent self-occlusion and complex movement. FLAG3D++ provides precise ground truth for such scenarios, offering a potential avenue for enhancing the performance of contemporary methods in addressing intricate postures. 2) Notably, prevailing skeleton-based action recognition [36], [37], [38] techniques demonstrate commendable performance when applied to previous benchmarks. However, their efficacy substantially diminishes when confronted with FLAG3D++ owing to its systematic categories. Therefore, the FLAG3D++ can server as a novel challenging benchmark. Our main contributions are summarized as follows:

1) We have presented a new benchmark named FLAG3D++ with versatile annotations of temporal intervals and action quality, highly accurate and dense 3D poses, detailed language instruction, and diverse video resources, which could provide fruitful data resources for fine-grained activity comprehension and fitness applications.
2) In light of FLAG3D++, we have presented two tasks, L-RAC and L-AQA, which incorporate natural language into traditional "vision-only" tasks of repetition action counting and action quality assessment. These offer more flexibility for real-world applications, and experimental results highlight the challenges inherent in these scenarios.
3) We have introduced a new framework named HL-GCN to perform a better cross-modal alignment, which hierarchically incorporates language instruction information into skeleton-based node features extracted by multiple graph convolutional layers. This is a versatile framework that can serve as a general feature extraction backbone for action

sequences with language instruction information.

4) We have evaluated various methods on FLAG3D++ under four different tasks. Extensive experiments have clearly demonstrated the effectiveness of the proposed HL-GCN on L-RAC and L-AQA. Additional experiments have also verified the generalized capability of FLAG3D++ for fine-tuning or transfer studies in various tasks.

This paper is built upon our conference paper [52] and significantly extended in several aspects. 1) We have expanded the fine-grained annotations of the temporal segments of actions in the untrimmed long sequences and how well an action is performed according to the language instructions. Based on these, we present two new practical applications: L-RAC and L-AQA. 2) We have devised a Hierarchical Language-Guided Graph Convolutional Network (HL-GCN) to hierarchically incorporate language features with skeleton sequence features, so that the different linguistic concepts can be better preserved during the cross-modal fusion. 3) We have enriched a series of ablation studies and visualization results. More experimental comparisons and in-depth analysis on FLAG3D++ have been included to demonstrate the effectiveness of our proposed HL-GCN for downstream tasks.

2 Related Work

Table 1 presents a comparison of our FLAG3D++ with the related datasets. FLAG3D++ provides fine-grained annotations of the temporal segments of actions in the untrimmed long sequences and how well an action is performed according to the language instructions, which enables two new practical applications of language-guided repetition action counting and language-guided action quality assessment. Furthermore, we also provide useful labels for human action analysis: SMPL mesh, action labels, and 3D keypoints. We briefly review the multimodal action datasets. We also discuss the relevant datasets and methods of the four tasks we study in this paper. The related works are as follows:

Multimodal Action Datasets. The academic community has consistently demonstrated keen interest in multimodal tasks, leading to the development of numerous image-text and video-text multimodal datasets [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67]. In human action analysis specifically, many datasets incorporate multimodal annotations [21], [51], [52], [68], [69]. Parmar et al. [21] discover that action quality assessment models can better represent actions if they additionally learn action descriptions and comments, and therefore construct a diving assessment dataset that incorporates textual content. Du et al. [51] also find that utilizing semantic information (such as commentary) can help models better understand and evaluate figure skating movements, and construct a figure skating dataset incorporating textual content. Zhang et al. [68] aim to comprehensively evaluate actions from multiple perspectives (action descriptions, professional objective assessments, qualitative evaluations, etc.) using narrative language, re-annotating the MTL-AQA [21] and FineGym [70] datasets. In addition to text annotations, FLAG3D++ offers comprehensive fine-grained action segmentation labels and action quality assessment annotations.

Repetition Action Counting. The objective of the Repetition Action Counting (RAC) task is to ascertain the number of repetitions of a periodic action within a given video. Early methodologies primarily focus on frequency domain analysis [71], [72], [73], [74], matching approach [75], or autocorrelation method [76], [77], [78], [79]. However, these approaches are constrained by the assumption of stationarity, wherein the periodic action is presumed to possess a fixed cycle length. Consequently, their performance excels in stationary datasets like YTsegments but deteriorates in non-stationary video datasets such as QUVA [80]. To address this limitation, Zhang et al. [17] propose a context-aware and scale-insensitive framework, and collect a larger dataset to alleviate the scarcity of data in this field. Concurrently, Dwibedi et al. [19] also propose the Countix dataset and develop a network named RepNet, which leverages the temporal self-similarity matrix to predict the counts accurately. However, Countix [19] and UCFRep [17] only involve short videos, which impedes the model's generalization capability in real-world scenarios. Furthermore, these datasets assign a numerical value as labels to each video, undermining the interpretability of the algorithm. Therefore, Hu et al. [18] present a dataset named RepCount to encompass videos of varied lengths with fine-grained annotations and propose a new multi-scale temporal correlation encoding network based on transformers to cope with both high and low-frequency actions. Despite these efforts, the existing datasets [17], [18], [19], [80] still suffer from limitations in terms of data quantity and they are primarily single-modal. In this context, we present FLAG3D++ providing an ample collection of extended action sequences accompanied by detailed language information and more fine-grained annotations, and devise the new HL-GCN to facilitate the exploration for involving natural language to the task of conventional RAC, referred to as HL-RAC.

Action Quality Assessment. The assessment of action quality has garnered significant attention within the field of computer vision due to its broad range of applications. Most existing approaches tackle this task by regression algorithm predicting a quality score [23], [81], [82], [83]. Numerous datasets have been introduced for action quality assessment across various domains, including diving [20], [21], [23], [84], [85], figure skating [85], [86], gymnastics [20], [82], [84], basketball [87], and artistic swimming [22]. The existing ones only primarily offer annotations at a coarse-grained level, and the approaches employed so far only yield a relatively coarse overall score. Nonetheless, we argue that the future paradigm of AQA should be interpretable and deeply integrated into human life. In light of this, we present FLAG3D++ along with its accompanying novel pipeline, HL-GCN, which holds significant potential to advance research within the community. FLAG3D++ provides fine-grained standards for each action. For a set of actions, we have issued scores that evaluate the action from various aspects over different time periods. Similar to the repetition action counting task, HL-GCN can assist in exploring the integration of linguistic text into action quality assessment.

Human Action Recognition. As the foundation of video understanding, pursuing diverse datasets has never stopped in action recognition. Existing works have explored various modalities, such as RGB videos [7], [88], [88], optical flows [89], audio waves [90], and skeletons [30]. Among these modalities, skeleton data draws increasing attention because of its robustness to environmental noises and action-focusing nature. During the past few years, various network architectures have been exploited to model the spatio-temporal evolution of the skeleton sequences, such as different variants of RNNs [91], [92], [93], CNNs [94], [95] and GCNs [30], [31], [96], [97], [98], [99]. In terms of skeletal keypoint, current action recognition datasets can be divided into two classes: one is 2D keypoint datasets [30], [38] extracted by pose estimation methods [100], [101], [102], [103], and the other is 3D keypoint datasets [40], [44], [104], [105] collected

TABLE 1: Comparisons of FLAG3D++ with the relevant datasets. FLAG3D++ consists of 180K sequences (Seqs) of 60 fitness activity categories (Cats). It contains both low-level features, including 3D key points (K3D) and SMPL parameters, as well as high-level language annotation (LA) to instruct trainers, sharing merits of multiple resources from MoCap system in laboratory (Lab), synthetic (Syn.) data by rendering software and natural (Nat.) scenarios. We evaluate various tasks in this paper, including human action recognition (HAR), human mesh recovery (HMR), and human action generation (HAG), language-guided repetition action counting (L-RAC), and language-guided action quality assessment (L-AQA). K2D: 2D keypoints, AL: action class label, S: action score, TB: temporal boundary, RI: repetition interval.

Dataset	Subjs	Cats	Seqs	Frames	LA	K3D	SMPL	Resource	Annotation	Task
PoseTrack [39]	-	-	550	66K	×	×	×	Nat.	K2D	HPE
Human3.6M [40]	11	17	839	3.6M	×	✓	-	Lab	AL	HAR,HPE,HMR
CMU Panoptic [41]	8	5	65	594K	×	✓	-	Lab	-	HPE
MPI-INF-3DHP [32]	8	8	-	>1.3M	×	✓	-	Lab+Nat.	-	HPE,HMR
3DPW [42]	7	-	60	51k	×	×	✓	Nat.	-	HMR
ZJU-MoCap [43]	6	6	9	>1k	×	✓	✓	Lab	AL,K2D	HAR,HMR
NTU RGB+D 120 [44]	106	120	114k	-	×	✓	-	Lab	AL	HAR,HAG
HuMMan [45]	1000	500	400K	60M	×	✓	✓	Lab	AL	HAR,HMR
HumanML3D [46]	-	-	14K	-	✓	✓	✓	Lab	-	HAG
KIT-ML [47]	111	-	3911	-	✓	✓	-	Lab	-	HAG
HumanAct12 [48]	12	12	1191	90K	×	×	✓	Lab	AL	HAR,HAG
UESTC [49]	118	40	25K	> 5M	×	✓	-	Lab	AL	HAR,HAG
Motion-X [50]	-	-	96k	13.7M	✓	✓	✓	Lab+Nat.	-	HAG,HMR
AQA-7 [20]	-	7	1106	175K	×	×	×	Nat.	S	AQA
MTL-AQA [21]	-	1	1412	135K	✓	×	×	Nat.	S + TB	AQA
FineDiving [23]	-	52	3000	288K	×	×	×	Nat.	S + TB	AQA
LOGO [22]	-	12	200	90	×	×	×	Nat.	S + TB	AQA
OlympicFS [51]	-	4	200	-	✓	×	×	Nat.	S	AQA
UCFRep [17]	-	23	526	87K	×	×	×	Nat.	RI	RAC
Countix [19]	-	-	8757	1M	×	×	×	Nat.	RI	RAC
RepCount [18]	-	-	1451	1M	×	×	×	Nat.	RI	RAC
Fit3D [1]	13	37	-	> 3M	×	✓	✓	Lab	RI	HPE,RAC
EC3D [15]	4	3	362	-	×	✓	-	Lab	AL	HAR
Yoga-82 [16]	-	82	-	29K	×	×	×	Nat.	AL	HAR,HPE
FLAG3D [52]	10+10+4	60	180K	20M	✓	✓	✓	Lab+Syn.+Nat.	AL	HAR,HMR,HAG
FLAG3D++(Ours)	10+10+4	60	180K	20M	✓	✓	✓	Lab+Syn.+Nat.	AL, RI, S, TB	HAR, HMR, HAG, L-RAC, L-AQA

by sophisticated equipment. However, most existing datasets are limited to a single domain of natural scenes. FLAG3D++ takes a different step towards cross-domain action recognition between rendered videos and real-world scenario videos.

Human Mesh Recovery. Human mesh recovery obtains well-aligned and physically plausible mesh results that human models can parametrize, such as SMPL [28], SMPL-X [29], STAR [106] and GHUM [107]. Current methods take keypoints [29], [108], [109], images [110], [111], [112], videos [113], [114], [115] and point clouds [116], [117] as inputs to recover the parametric human model under optimization [108], [118], [119] or regression [105], [120], [121] paradigm. And there are also ground-truth SMPL parameters provided by human datasets. They are registered by marker-less multi-view MoCap [29], [32], or marker/sensor based Mocap [40], [122]. SMPL can also be fitted with the rendered human scan in synthesis datasets [123], [124]. Easily recovered human poses of existing datasets cause the performance of human mesh recovery algorithms not to be fairly evaluated [125], whereas FLAG3D++ provides human poses with heavy self-occlusion and large movements. The work most related to ours is HuMMan [45], which contains large-scale and comprehensive multi-modal resources captured in a single MoCap room. In comparison, FLAG3D++ is complementary with rendered and natural videos, as well as more detailed language instructions to describe the activity.

3 THE FLAG3D++ DATASET

3.1 Taxonomy

The first challenge to construct FLAG3D++ is establishing a systematic taxonomy to organize various fitness activities. In previous literature, most existing fitness datasets [1], [15], [16] mix up all the activities. We present a deeper hierarchical lexicon as shown in Fig. 2, which contains three levels from roots to leaves, including body parts, fitness activity, and language instructions.

1) *Body Part.* For the first level, we share our thoughts with HuMMan [45] which uses the driving muscles as basic categories. However, numerous fine-grained muscles exist in the human body, and one activity might be driven by different muscles. We follow the suggestions of our fitness training coaches and choose ten parts of the human body with rich muscles as *chest, back, shoulder, arm, neck, abdomen, waist, hip, leg* and *multiple parts*[2].

2) *Fitness Activity.* Sixty everyday fitness activities are selected for the second level, linked to the corresponding body parts of the first level. For example, the activity *"Squat With Alternate Knee Lift"* is associated with the quadriceps femoris muscle of the *"Leg"*. The full list of all activities is included in the Appendix.

3) *Language Instruction.* We compose the third level of the lexicon with a set of language descriptions from the guidance of the training coaches to instruct users to accomplish the fitness activity. As an example shown in Fig. 2, the fitness activity *"Squat With Alternate Knee Lift"* is detailed as *"Stand with feet slightly wider than shoulder-width apart, bend your elbows and put your hands in front of your chest. Flex your hips and squat until your thighs are parallel to the ground, and keep your knees in the same direction as your toes when squatting..."*. There are about 3 sentences and 57 words for each fitness activity on average.

2. Some activities are driven by muscles of various body parts.

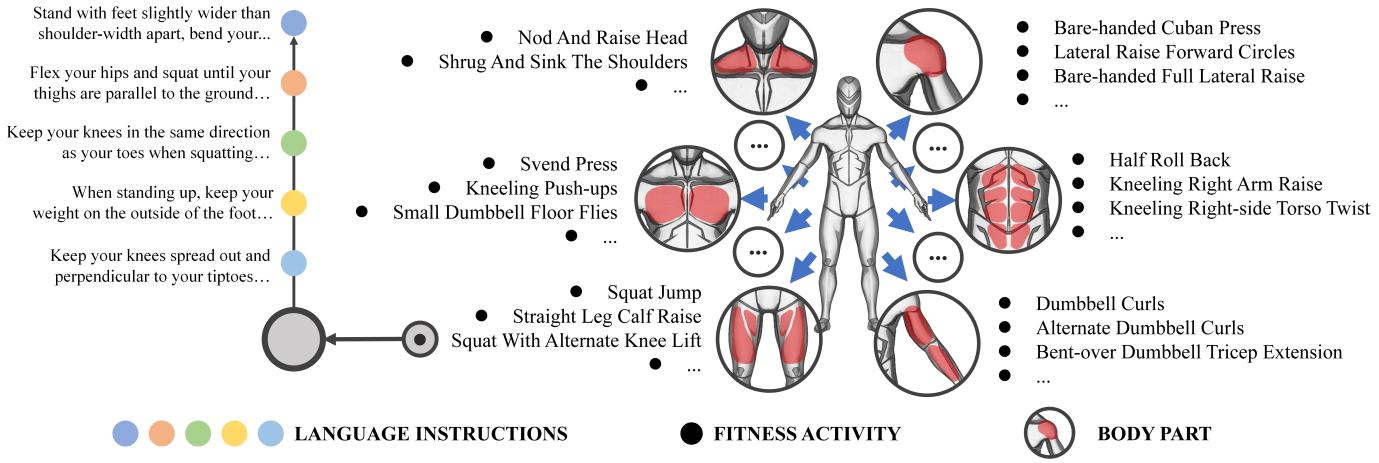


Fig. 2: An illustration of the taxonomy of our FLAG3D++ dataset, which is systematically organized in three levels as *body part*, *fitness activity* and *language instruction*. This figure details a concrete example of the *"Squat With Alternate Knee Lift"* activity that is mainly driven by the quadriceps femoris muscle of the *"Leg"*, while the corresponding language instructions are shown in the left.

3.2 Data Collection

We deploy high-precision MoCap equipment in an open lab to capture accurate human motion information. To obtain the rendered videos, we purchase kinds of virtual scenes and character models to make full use of the collected 3D skeleton sequences. In different environments, we record real-world natural videos. We agree with the volunteers and actively ensure that researchers can use these data. We further detail the data collection process below. More details can be found in the supplementary materials.

Data from MoCap System. Our MoCap system is equipped in a lab of 20 meters long, 8 meters wide, and 7 meters high. The lab uses the high-tech VICON [25] MoCap system to capture the actors' body part movements through optical motion capture. Cameras used in this system have a maximum resolution of 4096×4096. It is capable of 120fps while maintaining maximum resolution sampling. Ten volunteers perform the actions in the motion capture field. Through the professional machine, we can monitor the movements of volunteers in various forms, including masks, bones, and marker points. Moreover, we hire professional technicians to perform data restoration and motion retargeting based on high-precision original data so that we can ensure the accuracy and diversity of provided 3D motion data. Meanwhile, we ask each performer to wear MoCap clothes with 77 motion markers listed on Table A2 in supplementary materials. In total, we have 7200 motion sequences, where $7200 = 10(\textit{people}) \times 3(\textit{times}) \times 60(\textit{actions}) \times 4(\textit{motion retargeting})$.

During the data collection process of our previous conference version [52], to improve efficiency and ensure the standardization of data, we have established a set of standardized procedures. For most actions, we require volunteers to repeat the action eight times. However, the newly introduced L-RAC task requires the dataset to include a wider range of action repetition counts. To achieve this, we additionally capture approximately 20% of the skeleton-based sequences, which span a diverse range of repetition counts, from 2 to 62. This provides ample resources for training and testing in following repetition action counting task. We conduct a comparison of the statistical measures for the video duration and action repetition count in the dataset, specifically as shown in Table 2. The distribution of action repetition counts is shown in supplementary material.

Data from Rendering Software. To fully utilize the 3D human MoCap data, we use the rendering software Unity3D [126] to produce synthetic 2D videos with RGB color. For 2D videos, we purchase realistic scene models in Unity Asset Store [127], including indoor and outdoor scenes. As well, we predefine 6 camera positions in each scene. Our camera positions are dispersed around the avatar. However, we change parameters such as the focal length of each camera to ensure that the viewfinder fits and that the camera parameters are diverse. Specifically, we are appreciated that Renderpeople [128] provides several free character models. We select 4 avatars, import the skeleton information into avatars, and record the motion of the avatars in all directions. The resolution of these videos is 854 × 480, and fps is 30. Totally, we have 172,800 videos, where $172,800 = 7200(\textit{motion sequences}) \times 6(\textit{camera positions}) \times 4(\textit{virtual scenes})$.

Data from Real-world Environment. To obtain versatile data resources and add diverse scenes, we ask 10 extra people to record videos in different real-world scenarios. The recording process is executed using smartphones that can capture 1080p videos from the front view and side view simultaneously, to ensure the diversity of shooting angles. As well, before performing the activity, volunteers are asked to watch the instructional video and read the language instructions carefully. We have 7200 videos, where $7200 = 10(\textit{people}) \times 3(\textit{times}) \times 60(\textit{actions}) \times 2(\textit{views}) \times 2(\textit{scenes})$. And we have 24 subjects in all of the videos, where $24 = 10(\textit{mocap}) + 10(\textit{real-world}) + 4(\textit{render-people})$.

3.3 The Annotations

In order to facilitate FLAG3D++ for more fine-grained activity understanding tasks, we introduce versatile manual annotations based on the language instruction. To be specific, we select 1800 long action sequences from FLAG3D++ and annotate them for the applications of language-guided repetition action counting (L-RAC) and language-guided action quality assessment (L-AQA).

For L-RAC, the annotation process involves labeling the start and end frames of each periodic action within a given skeleton sequence, so that the repetitive times and occurrence interval can be grounded when taking the language instruction as a reference.

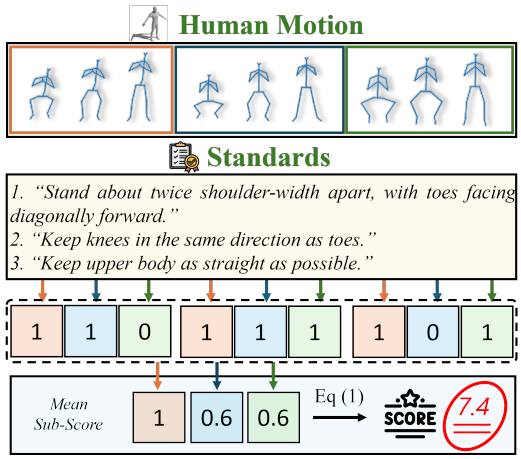


Fig. 3: The specific implementation of the annotation process. When presented with a skeleton sequence, the annotator marks every start frame and end frame for each repetition observed within the sequence. Moreover, for each repetition, three key scores are annotated simultaneously, which correspond to three action standards. Finally, all the key scores are summarized, and a final quality score is derived using the formula in Equation (1).

For L-AQA, we further annotate the quality of each action based on three well-established standards (instructions) from the hierarchical lexicon as depicted in Fig. 2. During the evaluation phase, each action i is assigned three key scores $[y_{i1}, y_{i2}, y_{i3}]$ with choices $[0, 1]$ by the annotator. The whole implementation of annotation is elucidated in Fig 3. Given the complexity of labeling segment points and assessing quality, employing conventional annotation tools would result in a substantial workload. To enhance annotation efficiency, we have developed a novel toolbox that expedites the labeling segment positions and scores. Fig. 4 shows an example interface, where the annotator can label the segment points and scores simultaneously. We engage a team of five workers who possess prior knowledge in the fitness domain. The annotation results provided by one worker undergo a rigorous verification and adjustment procedure conducted by another worker, ensuring the reliability and accuracy of the obtained annotations. Under this pipeline, the total duration required for the annotation process exceeds 90 hours. And the formula for computing the final score s is expressed mathematically as:

$$s = \delta + \sum_{j=1}^{3} g(\frac{1}{T}\sum_{i=1}^{T} y_{ij}), \tag{1}$$

where we define y_{ij} as the j-th score for clip i. T denotes the number of clips for a sequence while δ (set to 6) corresponds to the starting score which judges deduct points for errors and add points for requirements, and g is a linear function that maps scores with a range of $[0, 1]$ to scores with a range of $[-1, 1]$. The distribution of action scores is shown in supplementary material.

3.4 The Body Model

To facilitate different applications (*e.g.,* human mesh recovery), our FLAG3D++ adopts the SMPL [28] parametric model due to its ubiquity and generality in various downstream tasks. For more details, please refer to the supplementary materials.

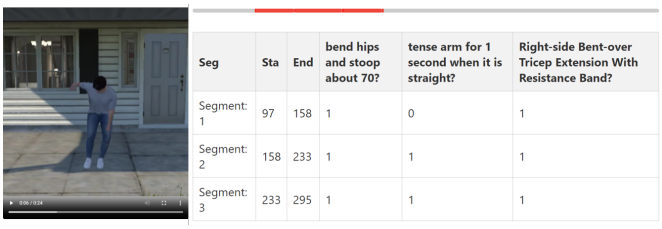


Fig. 4: The interface of our newly developed annotation tool.

TABLE 2: Dataset statistic of **Countix** [19], **UCFRep** [17], **RepCount [18]** and the proposed **FLAG3D++**. In comparison to other datasets, our dataset contains longer durations and a higher number of action repetitions with cleaner actions.

	UCFRep [17]	Countix [19]	RepCountA [18]	FLAG3D++
Duration Avg. ± Std.	8.15 ± 4.29	6.13 ± 3.08	30.67 ± 17.54	33.27 ± 16.16
Duration Min./Max	2.08/33.84	0.2/10.0	4.0/88.0	6.24/153.80
Count Avg. ± Std.	6.66	6.84± 6.76	14.99 ± 14.70	9.71 ± 7.99
Count Min./ Max	3/54	2/73	1/141	1/62

Specifically, the SMPL parameters comprise pose parameters $\theta \in \mathbb{R}^{N\times 72}$, shape parameters $\beta \in \mathbb{R}^{N\times 10}$ and translation parameters $t \in \mathbb{R}^{N\times 3}$, where N is the number of frames for each video. We obtain the SMPL parameters based on the captured keypoints and an optimization algorithm. In particular, the optimization process is composed of two stages, where the first stage is to get the shape parameter $\beta \in \mathbb{R}^{N\times 10}$, and the second stage is to gain the pose $\theta \in \mathbb{R}^{N\times 72}$ as well as translation parameter $t \in \mathbb{R}^{N\times 3}$. We denote the E_s and E_p as two objective functions for shape and pose optimization. In the first stage, the objective function is mathematically formulated as follows:

$$\begin{aligned} E_s(\beta) =& \frac{\lambda_1}{N}\sum_{(i,j)\in\mathcal{L}} ||\boldsymbol{J}_i(\mathbb{M}(\beta)) - \boldsymbol{J}_j(\mathbb{M}(\beta)) - \mathcal{P}(\boldsymbol{g}_i - \boldsymbol{g}_j)||_2^2 \\ &+\lambda_2||\beta||_2^2. \end{aligned} \tag{2}$$

Here $\boldsymbol{J_i}$ is the joint regressor for joint i, $\boldsymbol{g}$ is the ground truth skeleton, and $\mathbb{M}$ is the parametric model [28]. $\mathcal{L}$ and $\mathcal{J}$ represent the body limbs and joint sets, respectively. $\mathcal{P}$ is the projection function that projects the $g_i - g_j$ in the direction of $J_i - J_j$. Subsequently, the objective function in the second stage is:

$$E_p(\theta, t) = \lambda_3 \frac{1}{N}\sum_{j\in\mathcal{J}} \lambda_{p1}||\boldsymbol{J}_j(\mathbb{M}(\theta, t)) - \boldsymbol{g}_j||_2^2 + \lambda_4||\theta||_2^2. \tag{3}$$

In the equations above, different weights of $\lambda_k (k = 1, 2, 3, ...)$ are denoted for each loss term (see supplementary materials for details). We adopt the L-BFGS [129] method that satisfies the strong Wolfe conditions [130] for solving this optimization problem because of its memory and time efficiency.

4 Approach

In this part, we begin by introducing our Hierarchical Language-guided Graph Convolutional Network (HL-GCN) in Section 4.1. Then we present the utilization of this backbone in the language-guided repetition action counting in Section 4.2, and the language-guided action quality assessment in Section 4.3.

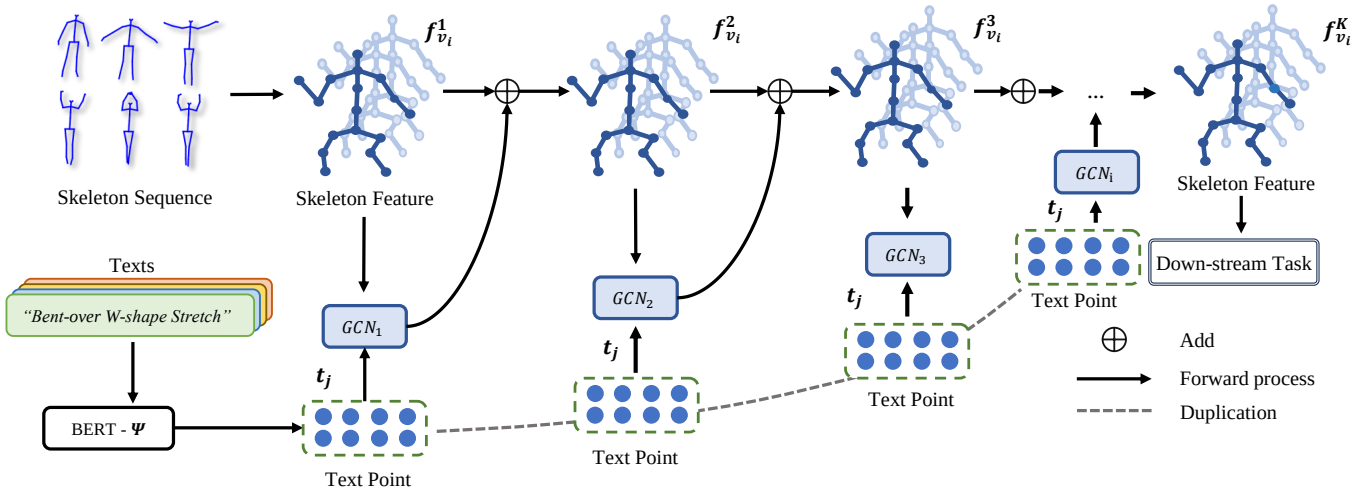


Fig. 5: Illustration of the Hierarchical Language-guided Graph Convolutional Network (HL-GCN). Compared with previous Spatial-Temporal Graph Convolutional Networks incorporating both spatial and temporal information to construct nodes in subsequent layers, our proposed method extends the paradigm to incorporate language information by integrating *text point* information. The *text point* information is obtained from a text encoder and serves as an attention control for each keypoint within the defined skeleton structure. For the detailed multimodal fusion process of each GCN layer, please refer to Figure 6.

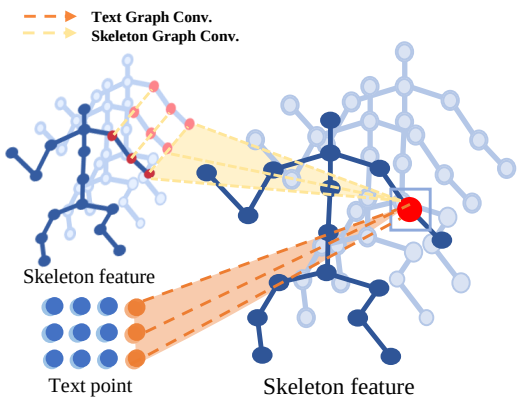


Fig. 6: The visual representation of the skeletal-text fusion graph convolution operation. The skeleton feature and text point are both treated as graph structure data, which are jointly processed through convolution operations and progressed to the next layer.

4.1 Language-guided Graph Convolutional Network

Human activity understanding encompasses various tasks, including action recognition, repetition action counting, quality assessment and many others, where the core techniques lie in effectively extracting discriminative representations from the input action sequences for the downstream tasks. When involving language instruction for reference, the key challenge lies in enhancing the original backbone to effectively accommodate multimodal information. Given that recent advances for skeleton-based video modeling predominantly rely on Graph Convolutional Networks (GCN) [30], [31], [97], [98], we extend the concept of graph convolution. We introduce a novel graph convolutional operation designed to not only fuse skeleton points but also concurrently integrate textual information, offering a dual-modality approach. The framework is illustrated in Fig. 5, referred to as HL-GCN.

f_v^k denotes the feature map of skeleton sequence in layer k, and following $f_{v_i}^{k+1}$ is the feature map for layer k and vertex i.

The Graph Convolutional Network (GCN) is a specialized architecture developed for processing graph-structured data, such as skeleton representations. In the field of skeleton-based action recognition, GCN has gained significant popularity and is widely regarded as one of the leading approaches. The primary workflow entails employing the graph convolution for the vertex in each layer. The graph convolution operation, applied to each vertex i in layer k, can be mathematically expressed as follows:

$$f_{v_i}^{k+1} = \sigma\Big(\sum_{v_j \in \mathcal{B}_{v_i}} \frac{1}{Z_{ij}} f_{v_j}^k \cdot w_{ij}^k\Big), \tag{4}$$

where σ represents the activation function simply combined with residual path. The sampling area of the convolution process for vertex v_i is denoted by $\mathcal{B}_{v_i}$. Within the graph convolution operation, w_{ij}^k signifies the weight function considering the mapping relationship between i and j, specifically devised by ST-GCN [30]. Additionally, Z_{ij} represents the normalizing term, which is equivalent to the cardinality of the corresponding subset.

In order to incorporate language information into the network, we introduce the *text point* attention mechanism to construct the adjacency matrix. This involves encoding the original text and projecting it into a text embedding, where the number of channels in the text embedding is as same as the number of skeleton keypoints, referred to as *text point*. Additionally, as shown in Fig. 5, in the input of each layer of GCN, we hierarchically introduce text information to ensure that text information consistently guides the refinement of features in the model. Specifically, these text points are utilized to construct the adjacency matrix combined with skeleton features. Within each layer of the HL-GCN framework, the graph convolution operation can be formulated as follows:

$$f_{v_i}^{k+1} = \sigma(f_a^k(v_i) + f_b^k(v_i) + f_b^k(v_i) * f_c^k(v_i, L)), \tag{5}$$

where "$*$" refers to the fusion operation and $f_a^{k+1}(v_i)$ is analogous with that in Equation (4):

$$f_a^{k+1}(v_i) = \sum_{v_j \in \mathcal{B}_{v_i}} \frac{1}{Z_{ij}} f_{v_j}^k \cdot w_{ij}^k. \tag{6}$$

The inclusion of the second component $f_b^k(v_i)$ serves to enhance the model's flexibility, enabling the attainment of an optimal structure during the training process rather than relying on a fixed structure. In this context, $\mathbb{V}$ represents the vertex set associated with the predefined graph, and a_{ij} denotes the learnable weight assigned to each pair of vertices. These learnable weights allow the model to adaptively determine the influence and connectivity between different vertex pairs, facilitating network's performance.

$$f_b^{k+1}(v_i) = \sum_{v_j \in \mathbb{V}} a_{ij}^k \cdot f_{v_j}^k \cdot w_{ij}^k. \tag{7}$$

The third component $f_c^k(v_i, L)$ involves performing the graph convolution operation on the text points t_j. Specifically, t_j is a general learnable weight assigned to the vertex v_j, indicating its significance within the context of the given language information.

$$f_c^{k+1}(v_i, L) = \sum_{v_j \in \mathbb{V}} t_j \cdot f_{v_j}^k \cdot w_{ij}^k, \tag{8}$$

$$t_j = (\text{norm}(\Psi(L)\theta_1)\theta_2)_j, \tag{9}$$

where Ψ is the text encoder to effectively process the language tokens L, and θ_1, θ_2 are text projectors in the form of MLP. The normalization method in the above formulation Equation (9) is the Layer Norm. We effectively fuse the language information $f_c^k(v_i, L)$ with learnable skeleton feature maps $f_b^k(v_i)$. We explore the fusion way of multiplication, cross attention and addition, and the results are shown in Table 3. Finally, we get the output features f_v^K after several GCN blocks.

4.2 Language-guided Repetition Action Counting

Given a sequence V comprising N frames, the objective of the conventional repetition action counting task is to accurately estimate the count of recurring actions within the sequence as $c = \mathcal{F}(V)$. Due to the sparse and wide distribution of the target domain in counting tasks, most existing works on repetition action counting do not directly learn an end-to-end mapping function $\mathcal{F}$. Instead, they initially generate predictions for immediate value $\mathcal{I}$ (*e.g.*, period length, period frequency) before subsequently further inference the number of repetitions as follows:

$$\mathcal{I} = \tau(\phi(V)), \tag{10}$$

where ϕ denotes the sequence encoder responsible for transforming the sequence into meaningful embeddings. And τ represents the prediction function employed to estimate the immediate value output $\mathcal{I} = [i_1, i_2, ..., i_N]$ from sequence embeddings. The immediate value $\mathcal{I}$ offers a range of potential options, including period length [19], density map [18], and salient pose [131]. Subsequently, the final count of repetitions is inferred based on the extracted immediate value $\mathcal{I}$ as $c = f(\mathcal{I})$, where the function f serves as the mapping from the extracted immediate value to the final count, and it typically comprises explicit rules without the need for learning, such as a division relation for period length.

To enhance the accuracy and generalizability of the counting model, leveraging our dataset's rich annotations, we propose the language-guided repetition action counting application (L-RAC). Considering our goal to incorporate linguistic information with skeletal data to obtain a language-related action representation to aid this task, we replace the previous encoder ϕ with HL-GCN:

$$f_v^K = \phi_{\text{HL-GCN}}(V, L), \tag{11}$$

Because periodic density map offers the spatial distribution beneficial to counting task [18], we follow TransRAC [18] to directly obtain the density maps as the immediate value $\mathcal{I}$ from the process $\mathcal{I} = \tau(f_v^K)$, where the similarity matrix and transformer block are utilized, as illustrated in the top part of Fig. 7. Given that the encoder employs HL-GCN for the extraction of sequence features, the entire method is henceforth referred to as HL-RAC.

We employ a data augmentation strategy to improve the distribution of the dataset. Specifically, for each action sequence, we randomly crop a portion of the repeated actions, making the number of repetitions diverse. Moreover, by applying this strategy, we effectively generate training and validation datasets that are 10 times larger than the original dataset.

4.3 Language-guided Action Quality Assessment

The objective of action quality assessment is to predict the score of a given action sequence. Given that our FLAG3D++ dataset offers detailed annotations and language instructions, we aim not only to predict the score of an action but also to pinpoint the precise areas of the body where performance is deficient and identify instances of poor execution among repetitions. This entails evaluating the quality of the entire action more finely and providing comprehensive explanations for observed shortcomings.

To precisely annotate each action in the action sequence, we need to separate each action, requiring the incorporation of the L-RAC before action quality assessment. Therefore, we first segment an action sequence V into different fragments $[C_1, C_2, ..., C_N]$.

$$[C_1, C_2, ..., C_N] = f_{\text{L-RAC}}[V, L], \tag{12}$$

where C_i corresponds to the i-th segment within $\mathcal{V}$, and $f_{\text{L-RAC}}$ denotes function aiming repetition action counting task. Subsequently, for each action segment we predict scores from three different criteria, and integrate all scores, as shown in Fig. 7.

To enable a model to predict according to different standards simultaneously and effectively enhance its generalizability, we propose a new application: language-guided action quality assessment (L-AQA). In this task, relevant action criteria are injected into the model as linguistic guidance, thereby specifically directing the model to generate scores related to those action standards:

$$\hat{s} = f_{\text{L-AQA}}[V(C_1, ..., C_N), S_1, S_2, S_3], \tag{13}$$

where S denotes the corresponding standard i and $f_{\text{L-AQA}}$ refers to the function to execute the action quality assessment task.

In the training stage, we utilize a single GT clip C_i, the criteria S_j, and the corresponding score y_{ij} for training. A learnable $\mathcal{G}$ predicts the score based on the criteria. $\mathcal{G}$ consists of the HL-GCN module and FC layers to predict whether it complies with standards. Based on this, $\mathcal{G}$ is actually a binary classifier that determines whether a given action clip C_i meets the standard S_j:

$$\hat{y_{ij}} = \mathcal{G}(C_i, S_j), \tag{14}$$

During the inference phase, we leverage action segments $[C_1, C_2, ..., C_N]$ predicted by the HL-RAC method. Finally, we calculate the overall score $\hat{s}$ by aggregating the key scores across all actions following the same logic with Equation (1).

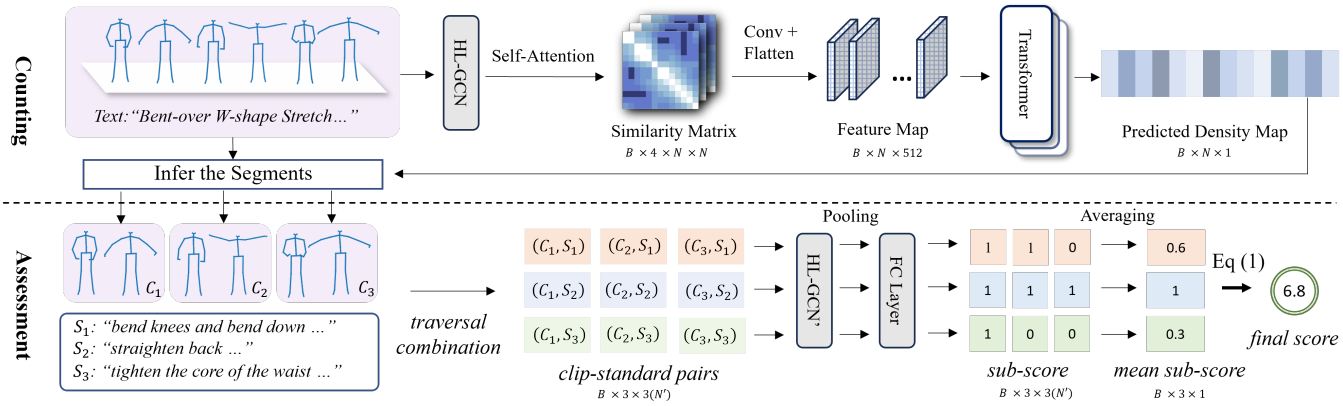


Fig. 7: Overview of our new paradigm for AQA. This picture comprises two distinct components. The upper section portrays the comprehensive pipeline of the HL-RAC framework, while the lower section illustrates the intricate process of generating the final score for the AQA. Initially, two separate networks are trained to execute the tasks of repetition action counting and key score prediction autonomously. Subsequently, upon completing the training phase, these networks are integrated to facilitate the inference of the AQA.

5 Experiments

In this section, we study four different tasks on our FLAG3D++ dataset as: (1) repetition action counting, (2) action quality assessment, (3) human mesh recovery, and (4) human action recognition. The first two tasks show the results of introducing natural language to the tasks of RAC and AQA, while the other verifies the capability of FLAG3D++ for various finetuning and transfer studies. The following describes the details of our experiments and results.

5.1 Repetition Action Counting

Fine-grained annotations for repetition action counting are included in FLAG3D++. To ensure proper evaluation, we systematically partition the dataset into train and test parts based on different actions. The performance of our proposed approach, HL-RAC, is assessed on the test set. We systematically record the corresponding metrics for actions in various body parts, as reported in Table 3. In order to evaluate the efficacy of the crucial components within our network, we conduct several ablations, the results of which are presented in Table 4. Furthermore, to assess the generalization capability of the proposed method, we conduct experiments on the RepCountA dataset [18]. We re-annotate the dataset to establish a correspondence between language and skeletal data. Specifically, we extract skeleton data from the RepCountA videos, filtering out inaccurate or misclassified data, and used the "type" field from the original annotations as the text input labels. We enrich these labels by carefully adding more detailed action descriptions. Since a large dataset is particularly essential for training an effective skeletal feature extraction backbone model, we apply the same data augmentation strategy as FLAG3D++ during the training phase. Apart from the results performed with the previous approach, we also evaluate several officially released models using video inputs. All results are shown in Table 7.

Experiment Setup. The dataset is split into train and test sets using a ratio of 50:10 for actions. And this selection of actions for each set is automatically performed randomly by the computer. Each action sequence is sampled to 128 frames. For videos longer than 128 frames, we use interval sampling to reduce them to 128 frames. For videos shorter than 128 frames, we pad them at the end. Two common metrics for repetition action counting are used to evaluate the performance of the model comprehensively.

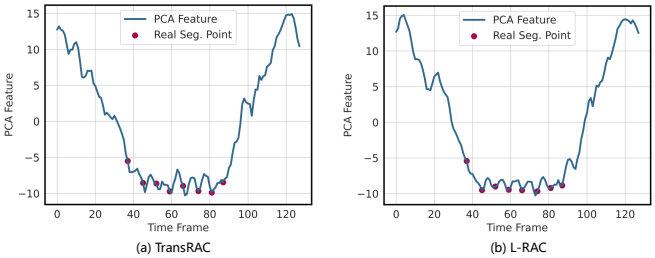


Fig. 8: 1D PCA projections of the HL-GCN encoder features over time. The troughs in features from the HL-GCN encoder are more fitting ground truth segments than TransRAC.

Mean Absolute Error (MAE). This metric actually calculates the relative error between the predicted count and the ground truth.
Off-By-One (OBO) count error. We closely follow [18] to define it as the accuracy for obtaining the right count.

$$\text{MAE} = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\frac{|c-\hat{c}|}{c}, \text{OBO} = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\mathbb{1}\{|c-\hat{c}|\leq 1\},$$

where $\mathcal{N}$ denotes the number of samples. c and $\hat{c}$ represents the ground truth counts and prediction counts, respectively.

Considering the leniency of the aforementioned metrics, we accordingly make adjustments to the threshold to further differentiate the performances of different models under evaluation. The extended off-by-one count errors are precisely defined as:

$$\text{OBO@}\alpha = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\mathbb{1}\{|c-\hat{c}|\leq \alpha\}, \text{OBO@Avg.} = \overline{\text{OBO@}\alpha_i},$$

Result and Analysis. We report the MAE and OBO@α in Table 3 where HL-RAC are compared with previous repetition action counting methods. Results show that our method exhibits the best performance. Thus incorporating the language information in repetition action counting task is meaningful. We also investigate three fundamental fusion functions and discovere that simple multiplication yields the optimal outcome in most cases. Table 4 highlights the crucial components in HL-RAC. The removal of our data

TABLE 3: Repetition action counting results on FLAG3D++ dataset. The subscripts "A, C, M" of HL-RAC represent the three different fusion methods of addition, cross-attention, and multiplication respectively, as stated in Section 4.1.

Method	Input	MAE ↓	OBO@0.4 ↑	OBO@0.6 ↑	OBO@0.8 ↑	OBO@1 ↑	OBO@Avg. ↑
Random	Skeleton	2.3630	0.0307	0.0307	0.0307	0.0955	0.0469
KNN	Skeleton	0.4220	0.1956	0.1956	0.1956	0.4446	0.2579
PCA + Peak Detect.	Skeleton	0.5464	0.1395	0.1395	0.1395	0.3137	0.1830
Fourier Analysis [132]	Skeleton	0.6769	0.0758	0.0758	0.0758	0.1827	0.1025
Wavelet Analysis [80]	Skeleton	0.6119	0.1182	0.1673	0.2067	0.2236	0.1790
RepNet [19]	Skeleton	0.2534	0.2907	0.4004	0.4598	0.5700	0.4302
TransRAC [18]	Skeleton	0.1375	0.4196	0.5273	0.6424	0.6921	0.5704
EasyShots [133]	Skeleton	0.3283	0.1443	0.2012	0.2587	0.3268	0.2328
IVAC [134]	Skeleton	0.3009	0.1759	0.2526	0.3171	0.4006	0.2866
DeTRC [135]	Skeleton	0.3385	0.2267	0.2267	0.2267	0.4984	0.2946
ME-RAC [136]	Skeleton	0.3220	0.2694	0.3400	0.4104	0.4575	0.3693
HL-RAC$_A$(Ours)	Skeleton + Lang.	0.1333	0.4320	0.5545	0.6305	0.6985	0.5789
HL-RAC$_C$(Ours)	Skeleton + Lang.	**0.1179**	0.4508	0.5614	0.6587	**0.7302**	0.6003
HL-RAC$_M$(Ours)	Skeleton + Lang.	0.1215	**0.4740**	**0.5899**	**0.6687**	0.7217	**0.6136**

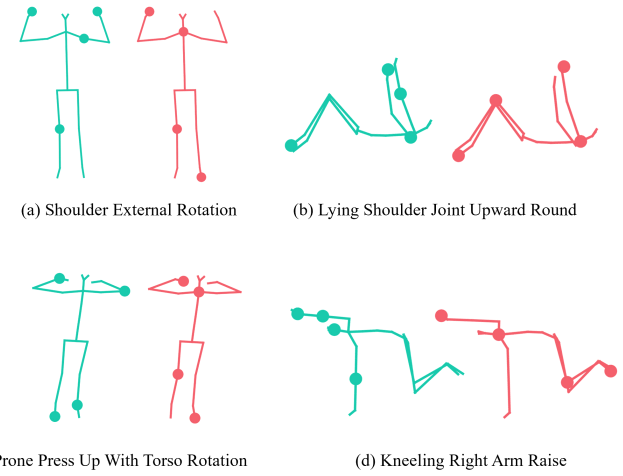
TABLE 4: Ablation studies on FLAG3D++ dataset for HL-RAC. It is evident that data augmentation plays a crucial role in enhancing the performance of models. Additionally, variations in other hyperparameters also influence the model's effectiveness.

	MAE ↓	OBO@1 ↑	OBO@Avg. ↑
(a) Data Augmentation			
w/o data augmentation	0.3007	0.3984	0.2760
w. data augmentation	**0.1179**	**0.7302**	**0.6003**
(b) Text Encoder in HL-GCN			
CLIP [137]	0.1458	0.6768	0.5592
BERT [138]	**0.1179**	**0.7302**	**0.6003**

augmentation strategy led to a significant decline in metrics, highlighting the effectiveness of this module. Additionally, variations in other hyperparameters also influence the model's effectiveness. Table 5 reports that the incorporation of language information has a positive impact on the model's performance across all categories of actions in the test set. The observed improvements in various action categories underscore the efficacy of leveraging language information to enhance the model's overall capabilities again. In Fig. 8, we present the visualizations of the graphs generated by HL-RAC and TransRAC. These visualizations highlight that the incorporation of text information in HL-RAC effectively enables the network to concentrate on specific keypoints, dependent on the action being performed. As a result, the model's ability to generalize across different actions is significantly enhanced.

Additionally, we explore the impact of varying levels of detail in the language captions on performance. Relevant experimental results are shown in Table 6. We use HL-RAC$_C$ as the baseline (Level 1). Level 1 contains the action's names, while Level 2 offers expanded descriptions of each action. Level 3 builds upon Level 2 by including professional fitness guidance, common mistakes, detailed movements, and the physical sensations involved. Please refer to the supplementary materials for examples of the different levels in language information. We find that Level 2 produces the best results, likely due to its more granular and accurate descriptions of the actions. Level 3 introduces excessive detail and irrelevant information (*e.g.*, sentences like "you may feel a slight burning sensation after many times" or "you will feel a distinct contraction"), causing a slight reduction in the performance.

Moreover, Table 7 presents the experimental outcomes on the RepCountA [18] dataset. The results clearly shows strong generalization of our approach and the effect of linguistic information on improving model's performance. We also observe that skeleton data input produces more reliable counting results than video modality input, which aligns with the findings of [131].



(a) Shoulder External Rotation (b) Lying Shoulder Joint Upward Round

(c) Prone Press Up With Torso Rotation (d) Kneeling Right Arm Raise

Fig. 9: Visualization results of graphs generated using HL-RAC with text information (represented in green) and TransRAC without text information (represented in red) for 4 actions from the test set. We highlight the top 4 points in the graph based on the weights significantly derived from the graph matrix. For the action *"Shoulder External Rotation,"* incorporating text information, our HL-RAC network exhibits enhanced attention towards the key point of the shoulder and hands, thus capturing its crucial movements more effectively. Similarly, in the case of *"Kneeling Right Arm Raise,"* our HL-RAC model demonstrates a more heightened focus on the right arm than TransRAC. For the actions *"Lying Shoulder Joint Upward Round,"* and *"Prone Press With Torso Rotation,"*, our model could effectively prioritize giving greater attention to the hands which are crucial keypoints in these movements.

5.2 Action Quality Assessment

As clearly shown in Fig 3, FLAG3D++ provides diverse levels of annotation information for the action quality assessment task. In this section, we make a comparison between our proposed

TABLE 5: Comparison of the HL-RAC and previous related repetition action counting methods on FLAG3D++ for actions from different body parts. Results that surpass all competing methods are **bold**. Our method achieves the best performance.

Method	Chest & Back		Shoulder		Arm		Neck & Abdomen		Waist		Hip		Leg	
	MAE ↓	OBO@1 ↑	MAE ↓	OBO@1 ↑	MAE ↓	OBO@1 ↑	MAE ↓	OBO@1 ↑	MAE ↓	OBO@1 ↑	MAE ↓	OBO@1 ↑	MAE ↓	OBO@1 ↑
Random	2.445	0.0944	2.5502	0.0891	2.6057	0.1090	2.4023	0.08833	2.2879	0.08596	2.1422	0.0850	2.3076	0.0900
KNN	0.3688	0.4598	0.4383	0.4347	0.2389	0.6378	0.2527	0.5600	0.5429	0.3929	0.4232	0.3566	0.6366	0.2933
PCA + Peak Detect.	0.5243	0.2818	0.5917	0.3905	0.6442	0.2242	0.5771	0.2383	0.5179	0.2666	0.5832	0.2800	0.3678	0.5016
Fourier Analysis [132]	0.6308	0.2133	0.6072	0.2210	0.6276	0.1893	0.6211	0.1400	0.6595	0.1745	0.6813	0.1683	1.0730	0.0950
Wavelet Analysis [80]	0.4662	0.4212	0.4578	0.3543	0.7048	0.0742	0.7035	0.1233	0.8049	0.0254	0.8085	0.0266	0.5177	0.3433
RepNet [19]	0.2129	0.6708	0.2498	0.6166	0.3147	0.5439	0.1380	0.7500	0.2847	0.4578	0.3134	0.3016	0.2760	0.5783
TransRAC [18]	0.1059	0.7622	0.1220	0.7166	0.0793	0.9015	0.0810	0.8783	0.1559	0.6070	0.2784	0.3849	0.1845	0.5400
EasyShots [133]	0.3670	0.2551	0.2637	0.3978	0.2676	0.3924	0.3057	0.3083	0.3279	0.4061	0.3312	0.2450	0.4816	0.1933
IVAC [134]	0.3069	0.3921	0.2737	0.3927	0.1813	0.6166	0.2539	0.3716	0.3893	0.2833	0.3311	0.4566	0.3314	0.3950
DeTRC [135]	0.3051	0.4960	0.3631	0.5036	0.2356	0.6666	0.3652	0.5799	0.3445	0.4219	0.3235	0.5133	0.4426	0.3549
ME-RAC [136]	0.3374	0.4464	0.2467	0.5463	0.3373	0.5257	0.2992	0.5166	0.3687	0.3964	0.3649	0.2966	0.3371	0.4199
HL-RAC$_A$(Ours)	0.1323	0.6866	0.1071	**0.7775**	0.0530	0.8833	0.0934	0.8716	0.1906	0.5333	0.1804	**0.5899**	0.1676	0.5883
HL-RAC$_C$(Ours)	**0.0909**	**0.7818**	0.1255	0.7355	0.0708	0.8848	0.1095	0.7350	**0.1266**	**0.7140**	**0.1623**	0.5749	**0.1571**	0.6200
HL-RAC$_M$(Ours)	0.1007	0.7015	**0.1015**	0.7557	**0.0492**	**0.9499**	**0.0704**	**0.8933**	0.1625	0.6140	0.2244	0.5299	0.1611	**0.6600**

TABLE 6: The effect of different granularities of text on the performance of the model.

Level Category	MAE ↓	OBO@0.4 ↑	OBO@0.6 ↑	OBO@0.8 ↑	OBO@1 ↑	OBO@Avg. ↑
Level 1: Action's name	0.1179	0.4508	0.5614	0.6587	0.7302	0.6003
Level 2: Action's details	**0.1094**	**0.4569**	**0.5928**	**0.6795**	**0.7460**	**0.6188**
Level 3: Action's instructions and suggestions	0.1234	0.4449	0.5547	0.6747	0.7391	0.6034

TABLE 7: Repetition action counting results on filtered RepCountA [18] dataset. The initial four rows in the first part of the table show results obtained by directly testing with the official pre-trained weights using video data as input. The subsequent rows display the test results of a model we train from scratch with skeleton-based input. Results that surpass all competing methods are bold.

Method	Input	MAE ↓	OBO@0.4 ↑	OBO@0.6 ↑	OBO@0.8 ↑	OBO@1 ↑	OBO@Avg. ↑
TransRAC [18]	Video	0.4518	0.0884	0.0884	0.0884	0.3451	0.1525
EasyShots [133]	Video	0.3347	0.1858	0.2566	0.3274	0.4159	0.2964
IVAC [134]	Video	0.4664	0.1061	0.1061	0.1061	0.2566	0.1438
Random	Skeleton	3.3590	0.0133	0.0133	0.0133	0.0433	0.0208
KNN	Skeleton	0.9549	0.0973	0.0973	0.0973	0.1681	0.1150
PCA + Peak Detect.	Skeleton	1.2306	0.0973	0.0973	0.0973	0.2831	0.1438
Fourier Analysis [132]	Skeleton	1.0108	0.0530	0.0530	0.0530	0.1061	0.0663
Wavelet Analysis [80]	Skeleton	0.6464	0.1504	0.2035	0.2300	0.2654	0.2123
RepNet [19]	Skeleton	0.8537	0.0088	0.0088	0.0088	0.0265	0.0132
TransRAC [18]	Skeleton	0.3186	0.1238	0.1858	0.2477	0.3008	0.2146
EasyShots [133]	Skeleton	0.3364	0.1238	0.2035	0.2477	0.2920	0.2168
IVAC [134]	Skeleton	0.3253	0.1238	0.1592	0.2035	0.3008	0.1969
DeTRC [135]	Skeleton	0.6137	0.0884	0.0884	0.0884	0.2477	0.1283
ME-RAC [136]	Skeleton	0.3074	0.1681	0.2743	0.3451	0.3805	0.2920
HL-RAC$_A$(Ours)	Skeleton + Lang.	0.2365	0.1769	0.2743	0.3451	0.4247	0.3053
HL-RAC$_C$(Ours)	Skeleton + Lang.	**0.2313**	0.1681	0.2477	0.3805	0.4513	0.3119
HL-RAC$_M$(Ours)	Skeleton + Lang.	0.2329	**0.2212**	**0.3097**	**0.3805**	**0.4601**	**0.3429**

approach and various conventional approaches employed in AQA using FLAG3D++ as the benchmark dataset.

Experiment Setup. The dataset is split into training and testing sets using 8:2 ratio for persons. And this selection of actions for each set is also performed randomly by the computer. The counting part of HL-AQA is also trained by this setup to ensure consistency throughout this task. Two common metrics for action quality assessment are used to evaluate the performance.

Spearman's rank correlation. Spearman's rank correlation is a statistical measure that quantifies the degree of monotonic relationship between two variables, indicating the tendency for one variable to increase or decrease as the other variable does.

$$\rho = \frac{\sum_{i=1}^{\mathcal{N}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{\mathcal{N}}(y_i - \bar{y})^2 \sum_{i=1}^{\mathcal{N}}(\hat{y}_i - \bar{\hat{y}})^2}},$$

where y and $\hat{y}$ are the rankings of two series, respectively. Moreover, we follow [20], [81] to calculate the Fisher's z-value as average performance across different classes.

Relative ℓ_2-distance. This metric is actually the normalized error between the predicted value and the ground truth value.

$$\text{R-}\ell_2 = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\left(\frac{|s_i - \hat{s}_i|}{s_{\max} - s_{\min}}\right).$$

Result and Analysis. Various methods are evaluated and their respective results are presented in Table 8. The outcomes indicate that our dataset poses significant challenges for the L-AQA task. Previous approaches have not only struggled to produce satisfactory results but also failed to yield finer-grained explanations. In contrast, our proposed method, HL-AQA, surpasses these approaches in metrics at most cases while providing fine-grained

TABLE 8: Comparison of our method and related action quality assessment approaches on FLAG3D++ for actions from different body parts. Results that surpass all competing methods are **bold**. Our method achieves promising performance in Spearman's rank correlation and comparable results in relative l_2-distance. Average correlation is computed by Fisher's z-value following [20], [81].

Sp. Corr. ↑	Chest & Back	Shoulder	Arm	Neck & Abdomen	Waist	Hip	Leg	Avg. Corr ↑
CoRe [81]	0.2612	0.1834	0.1487	0.3879	0.4723	**0.4291**	**0.4284**	0.3488
TSA [23]	0.2253	0.1683	0.1515	0.3788	0.4272	0.3871	0.3900	0.3181
Action-Net [82]	0.3967	0.4098	0.3628	**0.4574**	0.4130	0.4061	0.3864	0.4295
TPT [83]	0.2142	0.1692	0.1626	0.3882	0.4670	0.4027	0.4056	0.3322
Fine-parser [139]	0.1924	0.1731	0.1728	0.4036	**0.5172**	0.4055	0.4077	0.3439
HL-AQA(Ours)	**0.5336**	**0.5517**	**0.4627**	0.4048	0.3803	0.3860	0.3835	**0.4796**
R-ℓ_2 ↓	Chest & Back	Shoulder	Arm	Neck & Abdomen	Waist	Hip	Leg	Avg. R-ℓ_2 ↓
CoRe [81]	0.0960	0.1086	0.0974	0.0812	0.0774	0.0830	0.0806	0.0891
TSA [23]	0.1094	0.1219	0.1127	0.0899	0.0955	0.1031	0.1028	0.1050
Action-Net [82]	0.0826	0.0776	**0.0729**	**0.0714**	0.0841	0.0852	0.0829	0.0791
TPT [83]	0.0997	0.1202	0.1055	0.0883	0.0820	0.0926	0.0885	0.0967
Fine-parser [139]	0.0845	0.1007	0.0926	0.0756	**0.0718**	0.0896	0.0875	0.0860
HL-AQA(Ours)	**0.0602**	**0.0736**	0.0777	0.0826	0.0841	**0.0800**	**0.0800**	**0.0768**

TABLE 9: The impact of segmentation accuracy from counting algorithm on action quality assessment tasks across all categories.

Method	Sp. Corr. ↑	R-l_2 ↓
HL-AQA w. HL-RAC seg	0.4129	0.0792
HL-AQA w. GT seg	0.4485	0.0768

output. Nevertheless, for certain categories (such as the "waist", "hip" and "leg"), our method does not achieve the optimal performance in Spearman's rank correlation, whereas CoRE performs better. This is primarily due to two reasons: first, these categories have a more concentrated score distribution, making the ranking task more challenging. Thus contrastive learning methods tend to have an advantage, leading to better metrics. Second, these action categories are more complex, and the counting task tends to perform somewhat worse in these categories, which affects the performance of the action quality evaluation. Generally, our method demonstrates comparable results. When focusing solely on the relative R-ℓ_2 metric, we achieve state-of-the-art performance. However, our method does have some limitations. When the data distribution is relatively concentrated, making it difficult to distinguish subtle rankings among different actions, our method often fails to demonstrate significantly superior performance on Spearman's rank correlation. And our proposed method relies on the segmentation from the first-stage repetition action counting task. The more accurate the segmentation in the first stage, the better the final results of our proposed method. The impact of the counting component is shown in Tbale 9, where we replace counting predictions with ground truth values. With correct segmentation, the final predictions demonstrated increased accuracy.

5.3 Human Mesh Recovery

FLAG3D++ provides the SMPL [28] annotations, thus it is available to perform and evaluate popular methods for estimating 3D human poses and shapes. In this section, we first evaluate deep learning-based regression algorithms to verify that our dataset is qualified as a benchmark. Then we use the SMPL [28] annotation data to train ROMP [140] to improve its performance.

Experiment Setup. We select 300K frames for each scene and view during data selection to compose the subset for training and testing. In order to avoid potential continuity issues and information leakage (*e.g.*, two videos with the same action and human model but different repetitions are in different datasets), we select the first 20% videos for each scene and view them as the test set. We benchmark three typical methods: VIBE [141], BEV [142], and ROMP [140] on FLAG3D++ reporting MPJPE (mean per joint position error) and PA-MPJPE (Procrustes-aligned mean per joint position error) metrics. Results are presented in Table 10.

TABLE 10: Human mesh recovery accuracy on FLAG3D++ dataset. "↓" indicates that the lower value is better. "ft" represents that we have fine-tuned this method on our trainset.

Method	MPJPE ↓	PA-MPJPE ↓
VIBE [141]	376.67	106.27
BEV [142]	382.77	117.62
ROMP [140]	379.44	100.48
ROMP-ft [140]	**114.73**	**62.29**

TABLE 11: Performance on two types of challenging cases.

	Hard actions		Hard views	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
w/o	260.918	132.574	490.248	111.654
w. ft-FLAG3D++	**119.109**	**81.179**	**131.001**	**75.428**

Result and Analysis. These methods without fine-tuning achieved unsatisfactory MPJPE and PAMPJPE on the dataset. VIBE [141] and ROMP [140] achieved the best MPJPE and PA-MPJPE, respectively. But these metrics are still high, indicating that there is still much room for improvement of 3D shape estimation methods on FLAG3D++ dataset. Therefore, FLAG3D++ can serve as a new benchmark for 3D pose and shape estimation tasks. Since ROMP [140] achieved the top-1 PA-MPJPE, we fine-tuned it on FLAG3D++ with HR-Net [102] backbone. ROMP [140] could handle challenging cases and reach better MPJPE and PA-MPJPE after being fine-tuned on our dataset. It indicates that our dataset could benefit 3D pose estimation approach to improve their performance. To verify our ideas, we also test videos involving challenging actions and views specifically selected as shown in Table 11. Both situations with self-occlusion problems can be effectively mitigated after fine-tuning on FLAG3D++.

TABLE 12: Action recognition accuracy on the different benchmarks dataset. **Bold** results represent the performance of the model pre-trained on FLAG3D++, which confirms that the knowledge from our dataset can be transferred to other benchmarks.

Method	FLAG3D++	FineGym	NTU60-XSub
ST-GCN [30]	69.9	91.4 / **92.0** (**+0.6**)	89.0 / **89.0** (**+0.0**)
2s-AGCN [31]	81.6	91.8 / **92.1** (**+0.3**)	89.7 / **91.0** (**+1.3**)
MS-G3D [97]	73.6	92.7 / **93.4** (**+0.7**)	92.2 / **92.3** (**+0.1**)
CTR-GCN [98]	77.2	92.9 / **93.5** (**+0.6**)	90.6 / **90.8** (**+0.2**)
PoseC3D [38]	79.9	95.4 / **95.8** (**+0.4**)	93.7 / **93.9** (**+0.2**)

5.4 Human Action Recognition

FLAG3D++ contains both RGB videos and 3D skeleton sequences, maintaining abundant resources for 2D and 3D skeleton-based action recognition. We report our results on FLAG3D++ dataset and two other datasets, demonstrating the difficulty of our dataset and how fine-tuning on our dataset contributes to performance improvements on other datasets.

Experiment Setup. For data selection, we select the rendered videos from the front and side view in one scene (7200×3) for training samples and take all 7200 real-world videos for testing. We evaluate five state-of-the-art methods as ST-GCN [30], 2s-AGCN [31], MS-G3D [97], CTR-GCN [98] and PoseC3D [38]. For a fair comparison, we only adopt the joint modality to report our results. Table 12 presents the compared results.

Result and Analysis. Regarding the experiments, the accuracy demonstrates that our dataset poses a greater challenge in action recognition compared to other datasets. On the widely used NTU RGB+D 60 [104] and 120 XSub benchmark [44], Top-1 accuracy achieves and 93.7% and 86.0% respectively with PoseC3D [38], but 79.9% only on FLAG3D++. Unlike NTU RGB+D, which has a large proportion of daily actions in the indoor environment, FLAG3D++ focuses more on the classification of fitness actions, which requires more attention to fine-grained action distinctions. We finetune the models (pre-trained on FLAG3D++) on FineGym and NTU60. In Table 12, pre-trained models achieved better performance (**in bold**), particularly on FineGym, which shares some common grounds with FLAG3D++ such as the fine-grained nature of sports. Promising results show that our FLAG3D++ dataset can transfer beneficial signals for pre-trained models to significantly boost the performance of models on other datasets.

6 Future Work

Finally, we discuss some several potential directions for future works based on our FLAG3D++ dataset.

(1) Activity analysis with different data modalities. Our dataset provides a diverse rich array of RGB videos, skeletons, meshes, and multiple modalities. Users can easily explore various action understanding tasks across different modalities, such as RGB-based, skeleton-based, or mesh-based tasks. It's also encouraged to study different approaches to effectively fuse these modalities.

(2) Language-guided human action analysis. Our approach represents an initial step, opening extensive opportunities for future exploration of textual information to advance fine-grained action understanding using the FLAG3D++ benchmark. The future holds new possibilities for users to integrate linguistic information with video-based or skeleton-based representations. In exploring methods for text information integration, another task to consider is the visual grounding task. The language instructions of FLAG3D++ involve the critical steps of specific body parts to accomplish an activity. Grounding these key phases with the corresponding spatial-temporal regions could better bridge the domain gap between linguistic and visual inputs.

(3) Human action generation. Detailed language instructions and precise 3D skeleton-based motion sequences with SMPL [28] annotations are included in FLAG3D++, which facilitate the application of language-guided human action generation. We have evaluated multiple approaches on the dataset, and the associated experimental results can be found in the supplementary materials. We look forward to future users effectively utilizing this dataset to train or test more generative models.

7 Conclusion

In this paper, we have proposed FLAG3D++, a large-scale comprehensive 3D fitness activity dataset that shares the merits clearly over previous datasets from various aspects, including fine-grained annotations, language descriptions, highly accurate skeletons, and diverse resources. Both qualitative and quantitative experimental results have shown that FLAG3D++ significantly poses new challenges for multiple tasks like cross-domain human action recognition and human mesh recovery. In addition, we have proposed the HL-GCN, which hierarchically and effectively integrates skeleton information from motion sequences with language information using graph convolutional methods. We have demonstrated the effectiveness of HL-GCN on two applications as L-RAC and L-AQA. We hope FLAG3D++ will promote future research and more applications on fine-grained activity comprehension with language for the community.

Acknowledgments. This work was sponsored in part by the National Natural Science Foundation of China (Grant No. 62206153, 62321005, 62336004, 62125603), Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No: ZDSYS20210623092001004).

References

[1] M. Fieraru, M. Zanfir, S. C. Pirlea, V. Olaru, and C. Sminchisescu, "Aifit: Automatic 3d human-interpretable feedback models for fitness training," in *CVPR*, 2021, pp. 9919–9928.

[2] P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman *et al.*, "Proteingym: large-scale benchmarks for protein fitness prediction and design," *NeurIPS*, vol. 36, 2024.

[3] B. R. Barricelli, E. Casiraghi, J. Gliozzo, A. Petrini, and S. Valtolina, "Human digital twin for fitness management," *Ieee Access*, vol. 8, pp. 26 637–26 664, 2020.

[4] V. M. M. Kercher, K. Kercher, P. Levy, T. Bennion, C. Alexander, P. C. Amaral, A. Batrakoulis, L. F. J. G. Chávez, P. Cortés-Almanzar, J. L. Haro *et al.*, "2023 fitness trends from around the globe," *ACSM's Health & Fitness Journal*, vol. 27, no. 1, pp. 19–30, 2023.

[5] A. L. Gibson, D. R. Wagner, and V. H. Heyward, *Advanced fitness assessment and exercise prescription*. Human kinetics, 2024.

[6] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Research in Computer Vision*, vol. 2, no. 11, pp. 1–7, 2012.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011, pp. 2556–2563.

[8] Y. Li, Y. Li, and N. Vasconcelos, "Resound: Towards action recognition without representation bias," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 513–528.

[9] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.

[10] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE ICCV*, 2013, pp. 3192–3199.

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *CVPR*, 2019, pp. 6202–6211.

[12] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 046–12 055.

[13] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 244–253.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[15] Z. Zhao, S. Kiciroglu, H. Vinzant, Y. Cheng, I. Katircioglu, M. Salzmann, and P. Fua, "3d pose based feedback for physical exercises," *arXiv preprint arXiv:2208.03257*, 2022.

[16] M. Verma, S. Kumawat, Y. Nakashima, and S. Raman, "Yoga-82: a new dataset for fine-grained classification of human poses," in *CVPRW*, 2020, pp. 1038–1039.

[17] H. Zhang, X. Xu, G. Han, and S. He, "Context-aware and scale-insensitive temporal repetition counting," in *CVPR*, 2020, pp. 670–678.

[18] H. Hu, S. Dong, Y. Zhao, D. Lian, Z. Li, and S. Gao, "Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting," in *CVPR*, 2022.

[19] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Counting out time: Class agnostic video repetition counting in the wild," in *CVPR*, 2020.

[20] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *WACV*. IEEE, 2019, pp. 1468–1476.

[21] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *CVPR*, 2019, pp. 304–313.

[22] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *CVPR*, 2023.

[23] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *CVPR*, 2022, pp. 2949–2958.

[24] P. Yu, Y. Zhao, C. Li, J. Yuan, and C. Chen, "Structure-aware human-action generation," *ECCV*, 2020.

[25] "Vicon," https://www.vicon.com/hardware/cameras.

[26] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, and X. Savatier, "A study of vicon system positioning performance," *Sensors*, vol. 17, no. 7, p. 1591, 2017.

[27] N. Goldfarb, A. Lewis, A. Tacescu, and G. S. Fischer, "Open source vicon toolkit for motion capture and gait analysis," *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106414, 2021.

[28] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *TOG*, vol. 34, no. 6, pp. 1–16, 2015.

[29] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019, pp. 10 975–10 985.

[30] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018, pp. 7444–7452.

[31] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 12 026–12 035.

[32] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017, pp. 506–516.

[33] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration," in *ICCV*, 2021, pp. 1749–1759.

[34] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3d whole-body mesh recovery with component aware transformer," in *CVPR*, 2023, pp. 21 159–21 168.

[35] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. C.-F. Lin, "Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach," in *ECCV*, 2020.

[36] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark," https://github.com/open-mmlab/mmaction2, 2020.

[37] H. Duan, J. Wang, K. Chen, and D. Lin, "Pyskl: Towards good practices for skeleton action recognition," 2022.

[38] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *CVPR*, 2022, pp. 2969–2978.

[39] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *CVPR*, 2018, pp. 5167–5176.

[40] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *TPAMI*, vol. 36, no. 7, pp. 1325–1339, 2013.

[41] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *ICCV*, 2015, pp. 3334–3342.

[42] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018, pp. 601–617.

[43] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *CVPR*, 2021, pp. 9054–9063.

[44] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *TPAMI*, vol. 42, no. 10, pp. 2684–2701, 2019.

[45] Z. Cai, D. Ren, A. Zeng, Z. Lin, T. Yu, W. Wang, X. Fan, Y. Gao, Y. Yu, L. Pan *et al.*, "Humman: Multi-modal 4d human dataset for versatile sensing and modeling," *arXiv preprint arXiv:2204.13686*, 2022.

[46] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *CVPR*, 2022, pp. 5152–5161.

[47] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big data*, vol. 4, no. 4, pp. 236–252, 2016.

[48] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *ACM MM*, 2020, pp. 2021–2029.

[49] Y. Ji, F. Xu, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng, "A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition," *arXiv preprint arXiv:1904.10681*, 2019.

[50] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," *NeurIPS*, 2023.

[51] Z. Du, D. He, X. Wang, and Q. Wang, "Learning semantics-guided representations for scoring figure skating," *IEEE Transactions on Multimedia*, 2023.

[52] Y. Tang, J. Liu, A. Liu, B. Yang, W. Dai, Y. Rao, J. Lu, J. Zhou, and X. Li, "Flag3d: A 3d fitness activity dataset with language instruction," in *CVPR*, 2023, pp. 22 106–22 117.

[53] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[54] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018, pp. 2556–2565.

[55] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021, pp. 3558–3568.

[56] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *CVPR*, 2022, pp. 17 980–17 989.

[57] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv preprint arXiv:2111.11431*, 2021.

[58] C. He, Z. Jin, C. Xu, J. Qiu, B. Wang, W. Li, H. Yan, J. Wang, and D. Lin, "Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models," *arXiv preprint arXiv:2308.10755*, 2023.

[59] C. He, W. Li, Z. Jin, W. Wang, C. Xu, and D. Lin, "Opendatalab: Empowering general artificial intelligence with open datasets," https://opendatalab.com, 2022.

[60] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019, pp. 2630–2640.

[61] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *NeurIPS*, vol. 34, pp. 23 634–23 651, 2021.

[62] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve: Neural script

knowledge through vision and language and sound," in *CVPR*, 2022, pp. 16 375–16 387.
[63] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *CVPR*, 2022, pp. 5036–5045.
[64] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021, pp. 1728–1738.
[65] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NIPS*, vol. 35, pp. 25 278–25 294, 2022.
[66] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2443–2449.
[67] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.
[68] S. Zhang, S. Bai, G. Chen, L. Chen, J. Lu, J. Wang, and Y. Tang, "Narrative action evaluation with prompt-guided multimodal interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 430–18 439.
[69] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Visual-semantic alignment temporal parsing for action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[70] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *CVPR*, 2020, pp. 2616–2625.
[71] E. Pogalin, A. W. Smeulders, and A. H. Thean, "Visual quasi-periodicity," in *CVPR*. IEEE, 2008, pp. 1–8.
[72] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis, "Cyclic motion detection for motion based recognition," *Pattern recognition*, vol. 27, no. 12, pp. 1591–1603, 1994.
[73] A. Briassouli and N. Ahuja, "Extraction and analysis of multiple periodic motions in video sequences," *TPAMI*, vol. 29, no. 7, pp. 1244–1261, 2007.
[74] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *TPAMI*, vol. 22, no. 8, pp. 781–796, 2000.
[75] I. Laptev, S. J. Belongie, P. Pérez, and J. Wills, "Periodic motion detection and segmentation via approximate sequence alignment," in *ICCV*, vol. 1. IEEE, 2005, pp. 816–823.
[76] O. Azy and N. Ahuja, "Segmentation of periodically moving objects," in *ICPR*. IEEE, 2008, pp. 1–4.
[77] X. Tong, L. Duan, C. Xu, Q. Tian, H. Lu, J. Wang, and J. S. Jin, "Periodicity detection of local motion," in *ICME*. IEEE, 2005, pp. 650–653.
[78] Y. Ren, B. Fan, W. Lin, X. Yang, H. Li, W. Li, and D. Liu, "An efficient framework for analyzing periodical activities in sports videos," in *Image and Signal Processing*, vol. 1. IEEE, 2011, pp. 502–506.
[79] G. Li, X. Han, W. Lin, and H. Wei, "Periodic motion detection with roi-based similarity measure and extrema-based reference selection," *Consumer Electronics*, vol. 58, no. 3, pp. 947–954, 2012.
[80] T. F. Runia, C. G. Snoek, and A. W. Smeulders, "Real-world repetition estimation by div, grad and curl," in *CVPR*, 2018, pp. 9009–9017.
[81] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *ICCV*, 2021, pp. 7919–7928.
[82] L.-A. Zeng, F.-T. Hong, W.-S. Zheng, Q.-Z. Yu, W. Zeng, Y.-W. Wang, and J.-H. Lai, "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2526–2534.
[83] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *ECCV*. Springer, 2022, pp. 422–438.
[84] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *CVPR*, 2017, pp. 20–28.
[85] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*. Springer, 2014, pp. 556–571.
[86] S. Liu, X. Liu, G. Huang, L. Feng, L. Hu, D. Jiang, A. Zhang, Y. Liu, and H. Qiao, "Fsd-10: a dataset for competitive sports content analysis," *arXiv preprint arXiv:2002.03312*, 2020.
[87] G. Bertasius, H. Soo Park, S. X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," in *ICCV*, 2017, pp. 2177–2185.
[88] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.
[89] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014, pp. 568–576.
[90] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.
[91] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015, pp. 1110–1118.
[92] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *AAAI*, 2017, pp. 4263–4270.
[93] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *ICCV*, 2017, pp. 2117–2126.
[94] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *CVPR*, 2018, pp. 7024–7033.
[95] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "Pa3d: Pose-action 3d machine for video recognition," in *CVPR*, 2019, pp. 7922–7931.
[96] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 3595–3603.
[97] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *CVPR*, 2020, pp. 143–152.
[98] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *ICCV*, 2021, pp. 13 359–13 368.
[99] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *CVPR*, 2018, pp. 5323–5332.
[100] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.
[101] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *ICCV*, 2017, pp. 2640–2649.
[102] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.
[103] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *CVPR*, 2018, pp. 7103–7112.
[104] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.
[105] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *3DV*, 2018, pp. 484–494.
[106] A. A. A. Osman, T. Bolkart, and M. J. Black, "STAR: sparse trained articulated human body regressor," in *ECCV*, 2020, pp. 598–613.
[107] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Ghum & ghuml: Generative 3d human shape and articulated pose models," in *CVPR*, 2020, pp. 6184–6193.
[108] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*, 2016, pp. 561–578.
[109] Y. Zhang, Z. Li, L. An, M. Li, T. Yu, and Y. Liu, "Lightweight multi-person total motion capture using sparse multi-view cameras," in *ICCV*, 2021, pp. 5560–5569.
[110] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *CVPR*, 2018, pp. 6047–6056.
[111] R. A. Guler and I. Kokkinos, "Holopose: Holistic 3d human reconstruction in-the-wild," in *CVPR*, 2019, pp. 10 884–10 894.
[112] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *ICCV*, 2021, pp. 11 127–11 137.
[113] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3d human pose and shape from a video," in *CVPR*, 2021, pp. 1964–1973.
[114] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *CVPR*, 2019, pp. 5614–5623.
[115] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei, "Human mesh recovery from monocular images via a skeleton-disentangled representation," in *ICCV*, 2019, pp. 5349–5358.

[116] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Combining implicit function learning and parametric models for 3d human reconstruction," in *ECCV*, 2020, pp. 311–329.
[117] F. Hong, L. Pan, Z. Cai, and Z. Liu, "Garment4d: Garment reconstruction from point cloud sequences," in *NeurIPS*, 2021, pp. 27 940–27 951.
[118] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints," in *CVPR*, 2018, pp. 2148–2157.
[119] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *CVPR*, 2017, pp. 6050–6059.
[120] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018, pp. 7122–7131.
[121] H. Tung, H. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," in *NeurIPS*, 2017, pp. 5236–5246.
[122] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *IJCV*, vol. 87, no. 1, pp. 4–27, 2010.
[123] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017, pp. 109–117.
[124] Z. Cai, M. Zhang, J. Ren, C. Wei, D. Ren, J. Li, Z. Lin, H. Zhao, S. Yi, L. Yang *et al.*, "Playing for 3d human recovery," *arXiv preprint arXiv:2110.07588*, 2021.
[125] H. E. Pang, Z. Cai, L. Yang, T. Zhang, and Z. Liu, "Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms," in *NeurIPS*, 2022.
[126] "Unity3d," https://unity.com/.
[127] "Unity asset store," https://assetstore.unity.com/.
[128] "Renderpeople," https://renderpeople.com/.
[129] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," pp. 503–528, 1989.
[130] P. Wolfe, "Convergence conditions for ascent methods," pp. 226–235, 1969.
[131] Z. Yao, X. Cheng, and Y. Zou, "Poserac: Pose saliency transformer for repetitive action counting," *arXiv preprint arXiv:2303.08450*, 2023.
[132] E. Ribnick and N. Papanikolopoulos, "View-invariant analysis of periodic motion," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 1903–1908.
[133] S. Sinha, A. Stergiou, and D. Damen, "Every shot counts: Using exemplars for repetition counting in videos," *arXiv preprint arXiv:2403.18074*, 2024.
[134] H. Wang, Z.-Q. Cheng, Y. Du, and L. Zhang, "Ivac-p2l: Enhancing video action counting through irregular repetition priors," *arXiv preprint arXiv:2403.11959*, 2024.
[135] Z. Li, X. Ma, Q. Shang, W. Zhu, H. Ci, Y. Qiao, and Y. Wang, "Efficient action counting with dynamic queries," *arXiv preprint arXiv:2403.01543*, 2024.
[136] Y. Qiu, L. Niu, and F. Sha, "Multipath 3d-conv encoder and temporal-sequence decision for repetitive-action counting," *Expert Systems with Applications*, vol. 249, p. 123760, 2024.
[137] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.
[138] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[139] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. Peng, "Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 628–14 637.
[140] T. Fan, K. V. Alwala, D. Xiang, W. Xu, T. Murphey, and M. Mukadam, "Revitalizing optimization for 3d human pose and shape estimation: a sparse constrained formulation," in *ICCV*, 2021, pp. 11 457–11 466.
[141] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *CVPR*, 2020, pp. 5253–5263.
[142] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3d people in depth," in *CVPR*, 2022, pp. 13 243–13 252.

Yansong Tang (Member, IEEE) received the BS and PhD degrees both from the Department of Automation, Tsinghua University, in 2015 and 2020, respectively. From 2020 to 2022, he served as a postdoctoral fellow with the Department of Engineering Science of the University of Oxford. He is currently a tenure-track Assistant Professor of Shenzhen International Graduate School, Tsinghua University. His research interests include computer vision, pattern recognition, and video processing. In recent years, he has authored more than 40 papers in top peer-reviewed journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, and CVPR. He is a member of the IEEE.

Aoyang Liu is currently a master student in Tsinghua Shenzhen International Graduate School, Tsinghua University. Before that, he received the B.Sc. degree in Statistics from Beijing Jiaotong University in 2022. His current research interest is computer vision, including video understanding and multimodal understanding.

Jinpeng Liu is currently a master student in Tsinghua Shenzhen International Graduate School, Tsinghua University. Before that, he received the B.Sc. degree in the School of Intelligent System Engineering from Sun Yat-sen University in 2022. His current research interest is computer vision, including motion, video, and multimodal understanding.

Shiyi Zhang received the BS degree from the Department of Automation, Tsinghua University, Beijing, China, in 2023. He is currently working toward the PhD degree at the Shenzhen International Graduate School, Tsinghua University. His current research interest lies in video understanding, video generation, and video processing. He has authored one scientific paper published in CVPR. He serves as a regular reviewer member for a number of conferences, e.g., CVPR, ICCV, NeurIPS, and so on.

Wenxun Dai received the BS degree from the Xidian University, China, in 2023. He is currently a Master's Student at Tsinghua Shenzhen International Graduate School, Tsinghua University, China. His current research interests include motion generation and video understanding.

Jie Zhou (M'01-SM'04) received the BS and MS degrees both from the Department of Mathematics, Nankai University, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision and pattern recognition. In recent years, he has authored more than 100 papers have been published in TPAMI, TIP and CVPR. He is an associate editor for TPAMI and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE and Fellow of the IAPR.

Xiu Li (Member, IEEE) received the Ph.D. degree in computer-integrated manufacturing from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2000. From 2000 to 2002, she has served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China, where she has served as an Associate Professor from 2003 to 2010. Since 2016, she has been a Full Professor with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. Her research interests are in the areas of data mining, deep learning, computer vision, and image processing.

Jiwen Lu (M'11-SM'15-F'23) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, China. His current research interests include computer vision and pattern recognition. He was/is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He serves as the Co-Editor-of-Chief for PRL, an Associate Editor for the TIP, TCSVT, the TBIOM, and Pattern Recognition. He also serves as the Program Co-Chair of IEEE FG'2023, VCIP'2022, AVSS'2021 and ICME'2020. He received the National Outstanding Youth Foundation of China Award. He is an IEEE Fellow and IAPR Fellow.