

Fine-Tuning Large Multimodal Models for Fitness Action Quality Assessment

Gaetano Dibenedetto
University of Bari Aldo Moro
Bari, Italy
gaetano.dibenedetto@uniba.it

Marco Polignano
University of Bari Aldo Moro
Bari, Italy
marco.polignano@uniba.it

Elio Musacchio
University of Bari Aldo Moro
Bari, Italy
elio.musacchio@uniba.it

Pasquale Lops
University of Bari Aldo Moro
Bari, Italy
pasquale.lops@uniba.it

Abstract

Action Quality Assessment (AQA) plays an important role in evaluating human performance in different domains, including fitness, sports, and healthcare. This work introduces a novel AQA approach by fine-tuning large multimodal models (LMMs) for personalized activity evaluation. We used the Fitness-AQA Dataset, which provides detailed annotations of exercise errors under realistic conditions, and we adapt the LLaVA-Video model, a state-of-the-art LMM comprising the Qwen2 large language model and the SigLIP vision encoder. We have implemented a customized data preparation pipeline that transforms video-based exercise annotations into a conversational format specific for fine-tuning. To our knowledge, this study is among the first to fine-tune LMMs for AQA tasks and the very first to explore activity evaluation in this context. The experimental evaluation shows that our model achieves results slightly lower than the baseline, even though it is able to generalize across multiple exercises. The full-reproducible code is available on GitHub <https://github.com/GaetanoDibenedetto/UMAP25>.

CCS Concepts

• Computing methodologies → Computer vision problems.

Keywords

Large Multimodal Models, Fitness, Action Quality Assessment,

ACM Reference Format:

Gaetano Dibenedetto, Elio Musacchio, Marco Polignano, and Pasquale Lops. 2025. Fine-Tuning Large Multimodal Models for Fitness Action Quality Assessment. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '25)*, June 16–19, 2025, New York City, NY, USA. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3708319.3733684>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
UMAP Adjunct '25, New York City, NY, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1399-6/25/06
<https://doi.org/10.1145/3708319.3733684>

1 Introduction and Related Works

Action Quality Assessment (AQA) is a research field aiming at evaluating the quality of human movements and actions in domains such as sports, fitness, rehabilitation, and professional training. Effective AQA systems have the potential to provide actionable insights for performance improvement, injury prevention, and skill development. Traditional AQA methodologies, mainly based on hand-crafted features [23], pose estimation models [4, 30], or task-specific deep learning architectures [13, 29]. These have been explored across a range of domains, including Olympic sports [5, 18, 21, 23, 26], musical performances [22], skill-based activities [6], each with distinct challenges. In healthcare, it aids physiotherapy and rehabilitation by assessing movements in vulnerable populations [7, 14]. **However, research in real-world exercise and workout evaluation remains limited.**

Workout form assessment presents unique challenges due to real-world environmental variability, occlusions and dynamic nature of exercises. Videos captured in uncontrolled gym environments exhibit significant variations in camera angles, lighting, clothing, and equipment-induced occlusions. Detecting subtle movement errors remains a challenge for conventional human pose estimators. Ogata et al. [19] introduced a dataset focused on back squat assessment, relying on pose estimation to extract motion features. This approach struggled to detect certain movement errors due to the absence of mid-spine keypoints in the pose representation. To address these limitations, the **Fitness-AQA dataset** [20] was introduced as the first large-scale dataset for real-world workout assessment, covering back squat, barbell row, and overhead press. The videos have significant variations in camera angles, lighting conditions, and occlusions, capturing the complexities of action assessment in naturalistic environments. To tackle these challenges, researchers proposed a methodology based on self-supervised representation learning by developing two approaches. The first approach, named Motion Disentangling (MD), employs domain knowledge-informed self-supervised methods to leverage motion information and diverse environmental factors for robust representation learning. This method uses contrastive learning to disentangle local from global motion by utilizing half-cycles of exercises as triplets. An R(2+1)D architecture serves as the backbone, with augmentations such as horizontal flipping, partial masking, translation, and rotation applied to the triplets. The second approach introduces a self-supervised pose contrastive learning method, named Cross-View Cross-Subject Pose

Contrastive learning (CVCSPC), which generates pose-sensitive representations across different views, subjects, and instances in videos. Contrastive learning triplets are formed by selecting images from synchronized videos of humans in comparable poses. Both MD and CVCSPC methods require additional unlabeled data to effectively learn representations through contrastive learning. Additionally, a custom YOLOv3 model [25] is employed to detect barbells/weights and synchronize the videos.

A subsequent study by Gallardo et al. [10] proposed GYMetric-Pose for the Fitness-AQA dataset, which comprises the following steps: (1) Pose estimation, performed to extract 2D keypoints from video or image, followed by 3D pose regression. (2) Graph representation extraction: a spatial-temporal skeleton graph is constructed from connected joints and a line graph based on the skeleton graph, where nodes represent edge features and edges capture angular information. (3) Geometric Representation Extraction, which leverages angular information to incorporate finer skeletal details. (4) Classification layer to produce a binary output representing the presence or absence of error during the exercise execution.

Recent advances in Large Multimodal Models (LMMs) have demonstrated exceptional capabilities in vision-language tasks by integrating frameworks such as CLIP [24] and adaptation modules [3, 17] with Large Language Models (LLMs). For instance, OpenFlamingo [1] enhances pretrained language encoder layers by integrating gated cross-attention dense blocks. InstructBLIP [3] builds on BLIP-2 [15] by adding vision-language instruction tuning. LLaVA-Video [32] is a series of advanced large video-language models that expand the capabilities of open models in understanding video content. The main architecture component is given by SigLIP [31] as vision encoder and Qwen2 [28] as LLM. A key strength of this model lies in *LLaVA-Video_{SlowFast}* [8, 27] which optimizes the balance between the number of video frames processed and the count of visual tokens, all within the constraints of the limited context window of the language model and GPU memory. The application of LMMs to tasks such as visual question answering for video quality assessment [11] is receiving growing attention [9, 12, 33]. In this work, we performed Visual Instruction Tuning [17] to an LMM to detect execution errors in fitness exercises. Unlike existing methods trained on a per-exercise basis, our approach generalizes across multiple exercises without requiring a predefined classification step. Given the imbalance in the training set, we propose training strategies to mitigate this issue and enhance error detection performance. To fine-tune LLaVA-Video effectively, we structured the dataset in a conversational style, incorporating temporal annotations to enable the task of temporal action localization—pinpointing execution errors for specific body parts within video sequences. **To the best of our knowledge, this study is among the first to explore fine-tuning a LMM specifically for AQA tasks and the very first to apply it in the context of personalized activity evaluation.**

2 Proposed Approach

2.1 Dataset

In this research, we utilize the Fitness-AQA dataset [20], which, to the best of our knowledge, is the first and only dataset in the domain of fitness assessment that includes realistic environments.

Table 1: Fitness-AQA Dataset Statistics

Exercise	Error	Type	Samples	% Erroneous
BackSquat	Knee Inward	Video	1,623	14.29%
	Knee Forward	Video	1,623	68.33%
	Shallow	Image	3,611	43.87%
OverheadPress	Elbow	Video	2,260	34.38%
	Knee	Video	2,260	25.49%
BarbellRow	Lumbar	Image	14,778	15.68%
	Torso-Angle	Image	17,030	9.21%

Table 2: Training set distribution for Squat and Overhead Press (OHP). Squat includes Knees Forward (KFE) and Knees Inward (KIE) errors, while OHP covers elbow and knee errors.

Squat Errors		Samples	OHP Errors		Samples
KFE	KIE		elbows	knees	
✓	✗	781	✓	✗	392
✗	✓	37	✗	✓	551
✓	✓	159	✓	✓	98
✗	✗	402	✗	✗	880

This dataset accounts for subtle errors and naturalistic occlusions, offering a significant advancement over previous datasets. The only comparable dataset is that introduced by Ogata et al. [19], which is limited to a single human subject who deliberately simulates exercise errors without the use of weights such as barbells or dumbbells. In contrast, Fitness-AQA includes exercise clips collected from publicly available video-sharing platforms, such as Instagram and YouTube. It includes three exercises: (1) Back Squat, (2) Barbell Row, and (3) Overhead (Shoulder) Press.

As shown in Table 1, the dataset includes both labeled and unlabeled data. We focus only on the labeled videos, incorporating validation data into training. Previous works [10, 20] have primarily used the dataset for error detection, albeit the presence of annotations in the form of timestamps – indicating when erroneous movements occur – makes it well-suited for Temporal Action Localization. This is a key task in video understanding that aims to identify the start and end points of action instances, while determining their corresponding action class [16].

2.2 Data Augmentation

To address the training set imbalance (see Table 2), we applied data augmentation techniques to expand the dataset, by adding new examples. Inspired by [10], we generate new videos by using horizontal flipping, color inversion, and rotation. Additionally, we also apply combinations of these techniques, i.e. horizontal flipping + color inversion and horizontal flipping + rotation. This approach ensures that each sample has the same quantity of data in both its original perspective and its horizontally flipped version. These augmentations were implemented using the OpenCV library.

Horizontal flipping, a simple yet effective augmentation technique, mirrors video frames along the vertical axis. This method may be considered particularly useful, as it introduces variations in the visual representation of exercises without altering the fundamental semantics or motion dynamics of the activity. By incorporating horizontally flipped videos, the model is exposed to a wider range of visual patterns and spatial configurations, reducing the

risk of overfitting to specific orientations or viewpoints in the original dataset. This process increases the size of the training data, effectively increasing the diversity of input samples.

To achieve effective learning on unbalanced data, we implemented two different training strategies: (1) **"Two-Step"** approach, consisting of two training phases: in the first step the model is trained on the entire original dataset, without any data augmentation or removal of samples, while in the second step, the number of instances for each error type is reduced to match the least common error type in the dataset, i.e. 37. These samples are used both in their original form and with horizontal flip augmentation. (2) **"Dynamic-Step"** approach, which employs a dynamic strategy in which no data is removed, and augmented samples are dynamically generated to reach the highest number of instances for each type of exercise, e.g. 781 for squat.

2.3 Build Dataset for Fine-tuning

To fine-tune the LLaVA models, a specific conversation template is required¹. The structure is similar to the following:

```
1 "human": "<image>\nExpected input."
2 "gpt": "output desired"
```

Given that our task is to identify specific errors in a single repetition of different exercises, and the Fitness-AQA dataset provides annotations for up to two specific errors per video (see Table 1), we formatted our dataset for fine-tuning as following:

```
1 "human": "<image>\nThe subject is performing a Squat or a Overhead
   Press (OHP) exercise.
2 "gpt": "The subject is performing a Squat. The body parts
   committing errors are the following: knees forward and knees
   inward."
```

It is important to note that the dataset includes up to two errors per exercise, which translates into three possible responses: (1) Exercises executed with no errors. (2) Exercises with one identified error. (3) Exercises with two identified errors. These possibilities were formalized into the following response templates:

```
1 (1) "The subject is performing a {exercise_type}. He performs it
   correctly, without any subtle errors."
2 (2) "The subject is performing a {exercise_type}. The body part
   committing errors is the following: {error_type}."
3 (3) "The subject is performing a {exercise_type}. The body parts
   committing errors are the following: {error_type} and {
   error_type}."
```

In the above templates, `{exercise_type}` represents the type of exercise being performed (i.e. Squat or Overhead Press), while `{error_type}` indicates the specific errors identified in the exercise (e.g., elbows or knees). These placeholders are replaced with the appropriate labels from the dataset. To further enhance the robustness of the model and prevent overfitting to a fixed sentence structure, we introduced ten different sentence variations for each response type. These variations preserve the meaning while altering the wording, ensuring the model learns to generalize better rather than memorizing a single phrasing pattern. Additionally, to enhance the model's capability and potentially activate more cognitive reasoning processes, we incorporated the task of temporal action localization. This allows us to specify when an error is committed within the video clip by including the time range

of the erroneous execution and its end (in seconds). For instance, responses include temporal details such as:

```
1 (1) "The subject is performing an OHP. The body part involved in
   the error is as follows: knees from frame time 2.8s to 3.6s."
2 (2) "The subject is performing a Squat. The following body part is
   responsible for the error: knees forward from frame time
   2.62s to 4.58s."
```

To generate a precise time range in the model's output, the model must also have information about the video length and the specific frame timestamps at which data is extracted. To achieve this, we leveraged the time instructions already provided by the LLaVA framework. An example of a time instruction is: *"The video lasts for 3.10 seconds, and 8 frames are uniformly sampled at 0.00s, 0.43s, 0.87s, 1.30s, 1.73s, 2.17s, 2.60s, and 3.07s"*. Table 3 presents an example of model inference, including both video sequences (frames) and text.

3 Experimental Evaluation

The goal of the experimental evaluation is to assess the effectiveness of our approach in detecting erroneous execution in fitness exercises. Specifically, we aim to evaluate the model's ability to (1) classify exercise types accurately, (2) identify erroneous execution, specifically identifying the erroneous body part, and (3) localize these errors temporally. In this way, we demonstrate how our method generalizes across multiple exercises, while keeping competitive performance compared to baselines.

Existing works on the Fitness-AQA dataset [10, 20] were not trained to handle multiple types of exercise simultaneously, hence there are no metrics for baselines which recognize and distinguish between different exercise types. The only comparable quantitative evaluation is for the detection of execution errors in specific body parts. **Our work is the first to introduce temporal action localization within this dataset**, making comparisons with previous approaches in this area impossible.

As baselines, we used the three methods in Parmar et al. [20], e.g. CVCSPC which relies solely on images; MD, which relies solely on videos, and a combination of CVCSPC and MD. Moreover, we used GYMetricPose [10], using 3D human pose estimation data.

The metrics adopted to assess the performance of the models are: F1-score to evaluate the erroneous detection for each body part, and mean average precision (mAP) for the temporal action localization task. As described in [2], mAP is computed as Intersection over Union (IoU): the prediction is successful if its value exceeded the threshold value of the temporal overlap (IoU) between the prediction and ground truth segment (we used 0.5 as threshold for IoU). The Average Precision (AP) for a specific class is computed as the sum of videos with a successful prediction in that class over all videos in the same class. mAP is obtained by the mean of the AP over all classes.

Once the dataset is built following the conversational style of the LLaVA-Video (Sect. 2.3), we fine-tuned the *lmms-lab/LLaVA-Video-7B-Qwen2* model on four NVIDIA A100 GPUs (64GB each). Additional details on hyperparameters and training configurations are available in the GitHub repository. For the sake of completeness, we tested the original LMM without fine-tuning, and the results show that it is accurate in distinguishing the different types of exercise (squat vs OHP), but it is not able to detect any erroneous executions. This limitation is likely due to the subtle nature of the

¹https://github.com/haotian-liu/LLaVA/blob/main/docs/Finetune_Custom_Data.md

Table 3: Example of an instruction prompt, showing both video representation and text. Note that the number of frames used in this example is 8 to ensure a clear visualization of both video representation and time instruction in the text prompt.


Input	
	The video lasts for 3.10 seconds, and 8 frames are uniformly sampled from it. These frames are located at 0.00s, 0.43s, 0.87s, 1.30s, 1.73s, 2.17s, 2.60s, 3.07s. Please answer the following questions related to this video. The subject is performing a Squat or a Overhead Press (OHP) exercise. Which one is he making? Is he making a mistake? If so, what mistake is he making?
Output	The subject is performing a OHP. The body parts where errors are detected are as follows: elbows from frame time 0.0s to 2.78s and knees from frame time 0.0s to 0.89s.

Table 4: Performance comparison of erroneous detection in exercises. The table presents F1-score erroneous detection, for each body part, and mean average precision (mAP) for the temporal action localization task.

Method	Modality	F1-Score				AP				mAP
		Squat		OHP		Squat		OHP		
		KIE	KFE	Elbow Err.	Knees Err.	KIE	KFE	Elbow Err.	Knees Err.	
CVCSPC [20]	Image	0.5195	0.8286	0.4522	0.7203	–	–	–	–	–
MD [20]	Video	0.4186	0.8338	0.4552	0.8452	–	–	–	–	–
MD + CVCSPC [20]	Image, Video	0.5263	0.8468	–	–	–	–	–	–	–
GYMetricPose [10]	3D Pose	0.4398	0.8219	0.4175	0.8160	–	–	–	–	–
Dynamic-Step	Video+Text	0.0000	0.8155	0.3959	0.7866	0.0000	0.6475	0.1032	0.1622	0.2282
Two-Step	Video+Text	0.1955	0.6266	0.4575	0.7611	0.0410	0.5246	0.1534	0.1740	0.2232

errors occurring in the Fitness-AQA dataset, which often requires expert-level knowledge to be recognized. This demands to include additional knowledge in the model through fine-tuning. Detailed results are presented in Table 4. The performance of the model in detecting errors is comparable to the baselines, albeit slightly lower. It is worth to notice that baselines are trained separately for each exercise, whereas our approach generalizes across multiple exercises without requiring distinct models for each type. This makes our method more flexible and practical for real-world applications. The fewer number of examples for specific errors, such as knees inward error for squat, leads to a less accurate model, that is not able to gain performance even using augmented data. This specific error is one of the most challenging to detect, due to its subtle visual cues and its rarity in real-world movement execution. Same pattern of results is observed for the task of temporal action localization, where our model tries to identify the precise time range in which an execution error occurs. This capability is crucial for fine-grained movement analysis and personalized feedback, albeit a deeper investigation is necessary to improve the results.

4 Conclusions, Limitations and Future Works

In this paper we have presented an approach based on fine-tuning of LMMs for Action Quality Assessment in the fitness context. Our models have competitive performance on the Fitness-AQA dataset, with two main advantages: (1) ability to generalize across multiple exercises with a single model and (2) ability to perform temporal action localization, to identify the start and end points in videos where a specific type of error occurs.

This is a preliminary work, but some limitations hold. Due to the high memory requirements of LMMs, we are limited to process a maximum of 32 frames per sample on multiple GPUs with 64GB of VRAM. Given that the dataset consists of video clips recorded at 30

frames per second, this reduction results in a loss of temporal details, which may affect the model’s ability to accurately capture subtle exercise errors. To address this challenge, future work will explore improvements in the data processing pipeline through alternative data augmentation strategies. One promising approach is a clip-reduction method tailored to fitness exercises, where the lifting and lowering phases of an exercise are processed separately to better focus on distinct motion patterns. Another clip-reduction strategy to be investigating is splitting full-length videos into smaller clips of N consecutive frames, where N corresponds to the maximum number of frames supported by available computational resources. This would ensure that no frames are lost within each sub-clip, allowing the model to analyze complete motion trajectories, while remaining within memory constraints.

Acknowledgments

The research is partially funded by PNRR - Mission 4 ("Education and research") – Component 2 ("From research to business"), Investment 3.3 ("Introduction of innovative doctorates that respond to the innovation needs of companies and promote the hiring of researchers by companies") D.M.n. 117/2023 - CUP: H91I23000170007 and is supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 – Partnerships extended to universities, research centers, companies, and research D.D. MUR n. 341 del 15.03.2022 – Next Generation EU (PE0000013 – "Future Artificial Intelligence Research – FAIR" - CUP: H97G22000210007).

We extend our sincere gratitude to Naps Lab S.r.l.s.² for their support and collaboration in the realisation of this research.

²www.napslab.it

References

- [1] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *CoRR* abs/2308.01390 (2023). <https://doi.org/10.48550/ARXIV.2308.01390> arXiv:2308.01390
- [2] AbdulRahman M. Baraka and Mohd Halim Mohd Noor. 2022. Weakly-supervised temporal action localization: a survey. *Neural Comput. Appl.* 34, 11 (2022), 8479–8499. <https://doi.org/10.1007/S00521-022-07102-X>
- [3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instruct-BLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html
- [4] Gaetano Dibeneditto, Stefanos Sotiriou, Marco Polignano, Giuseppe Cavallo, and Pasquale Lops. 2025. Comparing human pose estimation through deep learning approaches: An overview. *Computer Vision and Image Understanding* (2025), 104297. <https://doi.org/10.1016/j.cviu.2025.104297>
- [5] Linfeng Dong, Wei Wang, Yu Qiao, and Xiao Sun. 2024. LucidAction: A Hierarchical and Multi-model Dataset for Comprehensive Action Quality Assessment. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=jj5isUwL3r>
- [6] Hazel Dougherty, Walterio W. Mayol-Cuevas, and Dima Damen. 2019. The Pros and Cons: Rank-Aware Temporal Attention for Skill Determination in Long Videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 7862–7871. <https://doi.org/10.1109/CVPR.2019.00805>
- [7] Chen Du, Sarah Graham, Colin Depp, and Truong Nguyen. 2021. Assessing Physical Rehabilitation Exercises using Graph Convolutional Network with Self-supervised regularization. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 281–285. <https://doi.org/10.1109/EMBC46164.2021.9629569>
- [8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-Fast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 6201–6210. <https://doi.org/10.1109/ICCV.2019.00630>
- [9] Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhui Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *CoRR* abs/2405.21075 (2024). <https://doi.org/10.48550/ARXIV.2405.21075> arXiv:2405.21075
- [10] Ulises Gallardo, Fernando Caro, Elune Hernández, Ricardo Espinosa, and Gilberto Ochoa-Ruiz. 2024. GYMetricPose: A light-weight angle-based graph adaptation for action quality assessment. In *37th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2024, Guadalajara, Mexico, June 26-28, 2024*, Gilberto Ochoa-Ruiz, Enrico Grisan, Sharib Ali, Rosa Sicilia, Lucía Prieto Santamaría, Bridget Kane, Christian Daul, Gildardo Sánchez-Ante, and Alejandro Rodríguez González (Eds.). IEEE, 43–50. <https://doi.org/10.1109/CBMS61543.2024.00016>
- [11] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. 2024. VQA²: Visual Question Answering for Video Quality Assessment. arXiv:2411.03795 [cs.CV] <https://arxiv.org/abs/2411.03795>
- [12] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *CoRR* abs/2307.16125 (2023). <https://doi.org/10.48550/ARXIV.2307.16125> arXiv:2307.16125
- [13] Chengxian Li, Xichong Ling, and Siyu Xia. 2023. A Graph Convolutional Siamese Network for the Assessment and Recognition of Physical Rehabilitation Exercises. In *Artificial Neural Networks and Machine Learning - ICANN 2023: 32nd International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, September 26-29, 2023, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 14257)*, Lazaros Iliadis, Antonios Papaleonidas, Plamen Angelov, and Chrisina Jayne (Eds.). Springer, 229–240. https://doi.org/10.1007/978-3-031-44216-2_19
- [14] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. 2021. Improving the Movement Synchrony Estimation with Action Quality Assessment in Children Play Therapy (ICMI '21). Association for Computing Machinery, New York, NY, USA.
- [15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html>
- [16] Ronglu Li, Tianyi Zhang, and Rubo Zhang. 2024. Weakly supervised temporal action localization: a survey. *Human. Tools Appl.* 83, 32 (2024), 78361–78386. <https://doi.org/10.1007/S11042-024-18554-9>
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html
- [18] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. 2021. EAGLE-Eye: Extreme-pose Action Grader using detail bird's-Eye view. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. IEEE, 394–402. <https://doi.org/10.1109/WACV48630.2021.00044>
- [19] Ryoji Ogata, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. 2019. Temporal Distance Matrices for Squat Classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2533–2542. <https://doi.org/10.1109/CVPRW.2019.00309>
- [20] Paritosh Parmar, Amol Gharat, and Helge Rhodin. 2022. Domain Knowledge-Informed Self-supervised Representations for Workout Form Assessment. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVIII (Lecture Notes in Computer Science, Vol. 13698)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 105–123. https://doi.org/10.1007/978-3-031-19839-7_7
- [21] Paritosh Parmar and Brendan Tran Morris. 2017. Learning to Score Olympic Events. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 76–84. <https://doi.org/10.1109/CVPRW.2017.16>
- [22] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. 2021. Piano Skills Assessment. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSp)*. 1–5. <https://doi.org/10.1109/MMSp53017.2021.9733638>
- [23] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. 2014. Assessing the Quality of Actions. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 8694)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 556–571. https://doi.org/10.1007/978-3-319-10599-4_36
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [25] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *CoRR* abs/1804.02767 (2018). arXiv:1804.02767 <http://arxiv.org/abs/1804.02767>
- [26] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. 2020. Learning to Score Figure Skating Sport Videos. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2020), 4578–4590. <https://doi.org/10.1109/TCSVT.2019.2927118>
- [27] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. *CoRR* abs/2407.15841 (2024). <https://doi.org/10.48550/ARXIV.2407.15841> arXiv:2407.15841
- [28] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuyang Liu, Zeyu Cui, Zhenru Zhang, Zhihang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. *CoRR* abs/2407.10671 (2024). <https://doi.org/10.48550/ARXIV.2407.10671> arXiv:2407.10671
- [29] Long Yao, Qing Lei, Hongbo Zhang, Jixiang Du, and Shange Gao. 2023. A Contrastive Learning Network for Performance Metric and Assessment of Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), 3790–3802. <https://doi.org/10.1109/TNSRE.2023.3317411>
- [30] Bruce X. B. Yu, Yan Liu, Keith C. C. Chan, and Chang Wen Chen. 2024. EGCN++: A New Fusion Strategy for Ensemble Learning in Skeleton-Based Rehabilitation Exercise Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 9 (2024), 6471–6485. <https://doi.org/10.1109/TPAMI.2024.3378753>

- [31] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 11941–11952. <https://doi.org/10.1109/ICCV51070.2023.01100>
- [32] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video Instruction Tuning With Synthetic Data. *CoRR* abs/2410.02713 (2024). <https://doi.org/10.48550/ARXIV.2410.02713> arXiv:2410.02713
- [33] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *CoRR* abs/2406.04264 (2024). <https://doi.org/10.48550/ARXIV.2406.04264> arXiv:2406.04264