Funded by the European Union NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | FAIR Future Artificial Intelligence Research | UniBa | SWAP researchgroup | nepsLab

Fine-Tuning Large Multimodal Models for Fitness Action Quality Assessment

Gaetano Dibenedetto, Elio Musacchio, Marco Polignano and Pasquale Lops
University of Bari Aldo Moro, Italy, *name.surname@uniba.it*

Goal & Motivation

We aim to improve fitness exercise evaluation by fine-tuning Large Multimodal Models (LMMs), using real-world annotated data. Our goal is to assess whether LMMs can generalize across different types of exercises and accurately detect even small mistakes with minimal supervision.

LLaVA-Video

A key strength of LLaVA-Video lies in LLaVA-Video*SlowFast* 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 architecture is composed of SigLIP as vision encoder and Qwen2 as language model.

Fitness-AQA Dataset

Background

We use the Fitness-AQA dataset, the first dataset in **fitness assessment** to include **realistic** home-like **environments**. It captures **subtle movement errors** and **natural occlusions**, providing a more realistic benchmark than prior datasets.

Dataset Statistics

Exercise	Error	Type	Samples	% Erroneous
BackSquat	Knee Inward (KIE)	Video	1,623	14.29%
	Knee Forward (KFE)	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%

Squat Errors		Samples
KFE	KIE	
✓	✗	781
✗	✓	37
✓	✓	159
✗	✗	402

OHP Errors		Samples
elbows	knees	
✓	✗	392
✗	✓	551
✓	✓	98
✗	✗	880

Methodology

Fine-Tuning Strategies

- **Two-Step**: First, fine-tune on the full original dataset. Then, fine-tune on a balanced subset (min class = 37), with horizontal flips.
- **Dynamic-Step**: To balance the data, we increase the number of samples, using augmentation techniques, to match the largest class size (e.g. 781 for squat)

Prompt Construction

Prompt Type	Example
Basic LLaVA Template	"human": "<image>\nExpected input." "gpt": "output desired"
Exercise and Error Detection	"human": "<image>\nThe subject is performing a Squat or an Overhead Press (OHP)." "gpt": "The subject is performing a Squat. Errors: knees forward and knees inward."
Temporal Action Localization	(1) "The subject is performing an OHP. Error: knees, from 2.8s to 3.6s." (2) "The subject is performing a Squat. Error: knees forward, from 2.62s to 4.58s."

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.

Experimental Evaluation

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

Observations

- Our approach performs **slightly below the baseline** in terms of F1-Score, but shows **stronger generalization** across different exercise types.
- The **Dynamic-Step strategy** generally performs better due to dataset balancing via augmentation.
- The **subtle nature of some errors (e.g. KIE)** makes even expert human annotation difficult, affecting all models.

Key Insights

✅ LMMs like LLaVA-Video can be fine-tuned for Action Quality Assessment

✅ The approach is generalizable, unlike traditional models that perform well only on specific exercises.

✅ Prompt engineering + temporal annotation integration is an effective method to adapt existing AQA datasets.