AI, Spot Me! Multimodal Models in the Gym

Gaetano Dibenedetto*, Elio Musacchio, Marco Polignano and Pasquale Lops

University of Bari Aldo Moro - Department of Computer Science, Via Orabona 4, Bari, 70125, Italy

Abstract

Action Quality Assessment (AQA) is essential for analyzing human performance in domains such as fitness, sports, and healthcare. In this paper, we present a novel method for AQA by fine-tuning large multimodal models (LMMs) to enable personalized activity evaluation. The Fitness-AQA dataset offers rich, error-level annotations of physical exercises in realistic conditions. To address this challenging dataset, we adapt the LLaVA-Video model, which combines the Qwen2 language model with the SigLIP vision encoder. Our approach stands out by being one of the first to apply LMM fine-tuning to AQA, and the first to target personalized feedback in this context. While the fine-tuned model slightly underperforms the baseline in standard metrics, it demonstrates a key strength: robust generalization across a wide variety of exercises without requiring per-exercise retraining. For the full implementation details, refer to [1][1].

Keywords

Large Multimodal Models, Fitness, Action Quality Assessment, Temporal Action Localization

1. Introduction and Related Works

Action Quality Assessment (AQA) aims to evaluate the quality of human movements across diverse domains such as sports, fitness, rehabilitation, and professional training. Robust AQA systems provide actionable insights for improving performance, preventing injury, and guiding skill development.

Traditional AQA methodologies, mainly based on hand-crafted features [2], pose estimation models [3], or task-specific deep learning architectures [4]. We propose fine-tuning an LMM to explore AQA in real-world workout assessment, a setting rarely studied, especially under challenges like occlusions, dynamic movements, and varied gym environments.

Ogata et al. [5] introduced a back squat dataset using pose estimation and motion feature extraction but were limited by missing keypoints in spinal regions. Subsequently, the **Fitness-AQA** dataset [6] was introduced, offering real-world recordings of back squats, barbell rows, and overhead presses under diverse conditions. Two contrastive learning strategies were proposed to learn robust representations:

- **Motion Disentangling (MD):** Uses self-supervised contrastive learning to disentangle local and global motion by organizing half-cycles of exercise as triplets, leveraging a R(2+1)D architecture and augmentations such as flipping, masking, and rotation.
- **CVCSPC:** Cross-View Cross-Subject Pose Contrastive learning generates representations by forming contrastive triplets from synchronized videos of similar poses across subjects and views.

Gallardo et al. [7] later proposed **GYMetricPose** for the Fitness-AQA dataset, comprising: (1) Pose estimation is performed to extract 2D keypoints from video or image input using VitPose [8], followed by 3D pose regression using MotionAGFormer [9]. (2) Two graph representations are created: a spatial-temporal skeleton graph 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 employs Edge Gated Convolutions to refine node and edge features

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.
gaetano.dibenedetto@uniba.it (G. Dibenedetto); elio.musacchio@uniba.it (E. Musacchio); marco.polignano@uniba.it (M. Polignano); pasquale.lops@uniba.it (P. Lops)
https://gaetanodibenedetto.github.io/ (G. Dibenedetto)
0000-0001-6083-3600 (G. Dibenedetto); 0009-0006-9670-9998 (E. Musacchio); 0000-0002-3939-0136 (M. Polignano); 0000-0002-6866-9451 (P. Lops)

for both graphs. This step leverages angular information to incorporate finer skeletal details. (4) A classification layer outputs a binary label indicating error presence in exercise execution.

Large Multimodal Models (LMMs) have recently emerged as powerful tools in vision-language tasks, integrating image encoders like CLIP [10] with large language models. Notable examples include OpenFlamingo [11], which integrated gated cross-attention dense blocks in the pretrained language encoder; InstructBLIP [12] enhances BLIP-2 with vision-language instruction tuning; LLaVA and LLaVA-Video [13] with the combination of Qwen2 [14] as a language model and SigLIP [15] as a vision encoder to process video via an efficient SlowFast sampling strategy [16].

Our Contribution. In this work, we perform **Visual Instruction Tuning** on LLaVA-Video for execution error detection in fitness exercises. Unlike prior models trained per exercise, our approach generalizes across multiple exercises without explicit classification. We introduce training strategies to handle class imbalance and reformat the dataset into a conversational structure with temporal annotations, enabling temporal localization of execution errors.

To the best of our knowledge, this is the first study to fine-tune a Large Multimodal Model for AQA and the first to explore personalized activity evaluation in this context. For the full implementation details, refer to [1][1].

2. Proposed Approach

2.1. Dataset

This study employs the **Fitness-AQA** dataset [6], the first dataset designed for real-world fitness assessment. Unlike prior datasets, it captures human exercises in unconstrained environments, including naturally occurring occlusions, varied lighting, and camera angles.

Fitness-AQA comprises video clips collected from public video-sharing platforms such as *Instagram* and *YouTube*, encompassing three commonly performed strength exercises: (1) Back Squat, (2) Barbell Row, and (3) Overhead (Shoulder) Press.

As shown in Table 1 in [1], the dataset contains both labeled and unlabeled samples. For our experiments, we focus only on the labeled videos, incorporating validation data into training.

While prior works [6, 7] focus on binary error classification, the dataset also includes temporal annotations marking the start and end of specific execution errors. These fine-grained labels make it well-suited for **Temporal Action Localization**, a video understanding task involving action boundary detection and classification over time [17]. This is essential for applications needing precise feedback on when and where errors occur during exercise.

2.2. Data Augmentation

To mitigate the class imbalance present in the training set (see Table 2 in [1]), we applied a series of data augmentation techniques aimed at expanding the dataset and enhancing its diversity. Inspired by the work in [7], we generated augmented video samples using the following transformations: horizontal flipping, color inversion, rotation, combined transformations i.e. horizontal flipping + color inversion and horizontal flipping + rotation.

Horizontal flipping, in particular, is a simple yet highly effective technique. By mirroring video frames along the vertical axis, it introduces spatial diversity without altering the semantic structure or motion dynamics of the activity. This increases the model's exposure to variations in orientation and viewpoint, reducing overfitting and improving its ability to generalize to unseen scenarios.

To further address data imbalance, we explore two distinct training strategies:

1. ***Two-Step***: This strategy involves a sequential training process. In the first phase, the model is trained on the full original dataset without any augmentation or data filtering. In the second phase, the

[1]accepted at ACM UMAP 2025 as a LBR: "Fine-Tuning Large Multimodal Models for Fitness Action Quality Assessment"

dataset is balanced by reducing the number of samples for each error class to match that of the least represented class (e.g., 37 instances). These samples are used both in their original form and with horizontal flip augmentation.

2. ***Dynamic-Step***: This strategy avoids any data removal. Instead, augmented samples are dynamically generated to increase the number of instances for each error type up to the maximum observed in the dataset (e.g., 781 for the squat class). This approach ensures a fully balanced training distribution while preserving the original dataset integrity.

2.3. Building the Dataset for Fine-Tuning

Fine-tuning the LLaVA models requires a specific conversation-style input template[2]. The expected format consists of alternating prompts between a human user and the model response.

Given that our objective is to detect specific movement errors within a single repetition of various exercises, we adapted the Fitness-AQA dataset accordingly. This dataset provides frame-level annotations indicating up to two types of errors per video (see Table 1 in [1]). We reformatted the data to match the LLaVA fine-tuning format, as illustrated in the following example:

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

Each exercise instance in the dataset can result in one of three possible annotation types: (1) Executed correctly, with no errors; (2) Performed with one error; (3) Performed with two distinct errors.

These possibilities are encoded using the following response templates:

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

`{exercise_type}` corresponds to the exercise being performed (e.g., Squat or Overhead Press), while `{error_type}` refers to the specific error (e.g., *knees forward*, *elbows*).

Furthermore, to enable fine-grained reasoning and strengthen the model's temporal understanding, we integrated a temporal action localization component. This involves specifying the time interval in which each error occurs within the video clip. For example:

```
"The subject is performing an OHP. The body part involved in the error is as follows: knees from frame time 2.8
    s to 3.6s."
"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 support this functionality, we provided the model with temporal data for each video, specifying both the total duration and the timestamps of the sampled frames. This informations follows the LLaVA time instruction format, such as:

> *"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 1 presents a full example of a training prompt, including both the video frames and corresponding textual descriptions used during model inference.

3. Experimental Evaluation

Our experimental evaluation assesses the effectiveness of the proposed approach in detecting execution errors during fitness exercises. We evaluate the model's ability to: (1) classify exercise type, (2) detect errors and affected body parts, and (3) localize errors temporally. Results show that our method generalizes across exercises while maintaining performance competitive with existing baselines.

[2]https://github.com/haotian-liu/LLaVA/blob/main/docs/Finetune_Custom_Data.md

Table 1
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 2
Performance comparison of erroneous detection in exercises. The table presents F1-score error 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 [6]	Image	0.5195	0.8286	0.4522	0.7203	–	–	–	–	–
MD [6]	Video	0.4186	0.8338	0.4552	0.8452	–	–	–	–	–
MD + CVCSPC [6]	Image, Video	0.5263	0.8468	–	–	–	–	–	–	–
GYMetricPose [7]	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

Existing approaches using the Fitness-AQA dataset [6, 7] focus on error detection within isolated exercises, without supporting multiple exercise types. No baselines exist for multi-exercise classification; only error detection on specific body parts is comparable. **To our knowledge, this is the first work to introduce temporal action localization in Fitness-AQA**, limiting comparisons in this area.

We evaluate our model against the following baselines proposed by Parmar et al. [6]: (1) **CVCSPC**, which operates on still images, (2) **MD**, which uses motion cues from video, and (3) a combination of both methods. Additionally, we include **GYMetricPose** [7], which uses 3D human pose estimation.

The evaluation metrics are as follows:

- **F1-score**, used to assess the detection of execution errors across body parts.
- **Mean Average Precision (mAP)**, used for the temporal action localization task, as defined in [18]. A prediction is considered correct if the Intersection over Union (IoU) between the predicted and ground-truth time segments exceeds a threshold (set to 0.5). The Average Precision (AP) is computed per class, and mAP is calculated as the mean of AP values across all classes.

For completeness, we also evaluated the pre-trained LMM without fine-tuning. While the model successfully distinguished between Squat and Overhead Press exercises, it failed to identify any execution errors. This limitation is likely due to the subtle nature of the annotated errors in the Fitness-AQA dataset, which often require domain expertise to detect.

Detailed performance comparisons are presented in Table 2. While our model shows slightly lower performance in error detection compared to baselines, it is important to emphasize that baseline models were trained individually for each exercise. In contrast, our approach learns from and generalizes across multiple exercise types using a single unified model. This flexibility makes our solution more suitable for real-world scenarios where handling diverse exercises within a single framework is advantageous.

Certain error types, such as *knees inward* during squats, occur infrequently and present only subtle visual cues. As a result, even with data augmentation, our model struggles to achieve high accuracy for such rare cases. This class imbalance impacts learning and remains a challenging aspect of the problem.

A similar trend is observed in the temporal action localization task. Although our model attempts to identify the precise time window during which an error occurs, its performance requires further improvement. Despite this, temporal localization remains a critical capability for detailed movement analysis and personalized feedback systems. Future work will focus on refining this aspect through improved temporal modeling and training strategies.

4. Conclusions, Limitations, and Future Work

In this paper, we presented a novel approach for Action Quality Assessment in the fitness domain, leveraging fine-tuned Large Multimodal Models (LMMs). Our method demonstrates competitive performance on the Fitness-AQA dataset, introducing two key contributions: (1) The ability to generalize across multiple types of exercises using a single unified model; (2) The capability to perform temporal action localization, identifying the start and end points in a video where specific execution errors occur.

While our results are promising, this work remains preliminary and is subject to several limitations. One major constraint is the substantial memory requirements of LMMs. Even with multiple GPUs equipped with 64GB of VRAM, we are limited to processing a maximum of 32 frames per video sample. Given that videos in the dataset are recorded at 30 frames per second, this frame reduction inevitably leads to the loss of fine-grained temporal details. Such limitations may impair the model's ability to detect subtle execution errors that unfold over short time intervals.

To overcome these challenges, future work will explore enhancements to the data processing pipeline. In particular, we aim to investigate alternative data augmentation strategies and temporal segmentation methods. One promising direction involves a clip-reduction strategy specifically tailored for fitness activities: splitting the exercise into separate clips corresponding to the lifting and lowering phases.

Another potential improvement involves dividing entire video sequences into smaller, non-overlapping clips of N consecutive frames, where N is defined by the model's maximum processing capacity. This would ensure full coverage of the motion trajectory without exceeding memory constraints, allowing for a more comprehensive temporal analysis across sub-clips.

Ultimately, addressing these limitations is key to building robust, scalable systems for automated fitness coaching and real-time movement analysis.

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. for their support and collaboration in the realisation of this research.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] G. Dibenedetto, E. Musacchio, M. Polignano, P. Lops, Fine-tuning large multimodal models for fitness action quality assessment, in: ACM UMAP Adjunct 2025, New York City, USA, June 16-19, 2025, 2025. doi:10.1145/3708319.3733684.

[2] H. Pirsiavash, C. Vondrick, A. Torralba, Assessing the quality of actions, in: D. J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI, volume 8694 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 556–571. doi:10.1007/978-3-319-10599-4_36.

[3] G. Dibenedetto, S. Sotiropoulos, M. Polignano, G. Cavallo, P. Lops, Comparing human pose estimation through deep learning approaches: An overview, Computer Vision and Image Understanding (2025) 104297. doi:https://doi.org/10.1016/j.cviu.2025.104297.

[4] L. Yao, Q. Lei, H. Zhang, J. Du, S. Gao, 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. doi:10.1109/TNSRE.2023.3317411.

[5] R. Ogata, E. Simo-Serra, S. Iizuka, H. Ishikawa, Temporal distance matrices for squat classification, in: IEEE Conference on CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 2533–2542. doi:10.1109/CVPRW.2019.00309.

[6] P. Parmar, A. Gharat, H. Rhodin, Domain knowledge-informed self-supervised representations for workout form assessment, in: S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), ECCV 2022, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVIII, Springer, 2022. doi:10.1007/978-3-031-19839-7_7.

[7] U. Gallardo, F. Caro, E. Hernández, R. Espinosa, G. Ochoa-Ruiz, Gymetricpose: A light-weight angle-based graph adaptation for action quality assessment, in: G. Ochoa-Ruiz, E. Grisan, S. Ali, R. Sicilia, L. P. Santamaría, B. Kane, C. Daul, G. Sánchez-Ante, A. R. González (Eds.), CBMS 2024, Guadalajara, Mexico, June 26-28, 2024, IEEE, 2024, pp. 43–50. doi:10.1109/CBMS61543.2024.00016.

[8] Y. Xu, J. Zhang, Q. Zhang, D. Tao, Vitpose: Simple vision transformer baselines for human pose estimation, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

[9] S. Mehraban, V. Adeli, B. Taati, Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network, in: IEEE/CVF WACV 2024, Waikoloa, HI, USA, January 3-8, 2024, IEEE, 2024, pp. 6905–6915. doi:10.1109/WACV57701.2024.00677.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[11] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Y. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, L. Schmidt, Openflamingo: An open-source framework for training large autoregressive vision-language models, CoRR abs/2308.01390 (2023). doi:10.48550/ARXIV.2308.01390. arXiv:2308.01390.

[12] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. C. H. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

[13] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, C. Li, Video instruction tuning with synthetic data, CoRR abs/2410.02713 (2024). doi:10.48550/ARXIV.2410.02713. arXiv:2410.02713.

[14] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, H. Wei, et al., Qwen2 technical report, CoRR abs/2407.10671 (2024). doi:10.48550/ARXIV.2407.10671. arXiv:2407.10671.

[15] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: IEEE/CVF, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, 2023, pp. 11941–11952.

[16] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: 2019 IEEE/CVF, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 6201–6210. doi:10.1109/ICCV.2019.00630.

[17] R. Li, T. Zhang, R. Zhang, Weakly supervised temporal action localization: a survey, Multim. Tools Appl. 83 (2024) 78361–78386. doi:10.1007/S11042-024-18554-9.

[18] A. M. Baraka, M. H. M. Noor, Weakly-supervised temporal action localization: a survey, Neural Comput. Appl. 34 (2022) 8479–8499. doi:10.1007/S00521-022-07102-X.