FIAS: Fitness Insight Analysis System. A Framework for Explainable Skeleton-based Action Recognition to Empower LLM Fitness Coaches

An Kai Chen
National Tsing Hua University
Hsinchu, Taiwan
kylechen.ankai@gmail.com

Chiu Yun Chang
hs102a4a04@gmail.com

Liong Zheng Ee
lze0603@gmail.com

Abstract

Skeleton-based action recognition is a cornerstone of modern fitness applications, yet it suffers from significant performance degradation when faced with novel viewpoints. This paper presents FIAS, a comprehensive framework for building a robust and, crucially, explainable fitness analysis system. We first conduct a systematic analysis of the ST-GCN++ model, targeting the challenges of cross-view generalization. Our experiments establish that a mixed-view training strategy combined with dynamic 2D modalities and augmentations is paramount, achieving a state-of-the-art offline accuracy of 99.53% within our dataset's distribution, conclusively outperforming 3D-lifted counterparts.

Beyond offline accuracy, we evaluated the system's viability for real-time deployment. Our temporal analysis shows the model achieves a strong category-level mean Temporal IoU (mIoU) of 0.478 with a rapid 0.670s average responsiveness, confirming its suitability for streaming applications. Finally, we introduce a novel XAI pipeline that utilizes Grad-CAM to visualize the model's biomechanical evidence, which is then translated into actionable, human-readable coaching advice by a Large Language Model (LLM). This work provides a clear directive for building practical fitness AI: prioritize a diverse, augmented 2D training pipeline and integrate explainability methods to transform a high-accuracy classifier into a trustworthy, real-time coaching tool.

1. Introduction

Human Action Recognition (HAR) has become a pivotal technology in a myriad of applications, from autonomous driving and surveillance to human-computer XAI interaction. Its recent application in automated fitness coaching has gained significant traction, promising to democratize personal training by providing real-time feedback. Skeleton-based methods are particularly well-suited for this domain due to their computational efficiency and robustness to environmental variations.

However, a significant and often underestimated challenge in deploying these systems in uncontrolled environments is viewpoint variance. A model trained on exercise data captured from a single perspective may exhibit a precipitous decline in performance when the user's orientation changes. This lack of cross-view generalization is a major barrier to creating reliable fitness applications. A common assumption is that lifting 2D skeletons to 3D will resolve this; however, 3D data estimated from a single 2D source can be noisy and carry inherent view-dependent biases, potentially failing to provide the expected robustness.

This paper addresses this challenge through a comprehensive and systematic experimental framework. We developed FIAS (Fitness Insight Analysis System) to explore the intricate relationship between input data dimensionality (2D vs. 3D), feature modality (joint, bone, motion), and training strategy (single-view vs. mixed-view). Our goal is to identify an optimal configuration for a skeleton-based model that is both highly accurate and robust to viewpoint changes. Our key contributions are threefold:

1. We provide an exhaustive cross-view generalization analysis, quantifying the severe performance degradation of single-view models.

2. We demonstrate that a mixed-view training strategy combined with simpler 2D dynamic features (bone and joint motion) decisively outperforms more complex 3D models.

3. We introduce a novel XAI framework that leverages Grad-CAM and an LLM to translate model predictions into explainable, human-centric coaching feedback, validating its efficacy with a qualitative case study.

Our results offer a clear framework for practitioners, emphasizing that data diversity and effective 2D feature representation are more critical than data dimensionality for building robust and practical fitness AI.

2. Related Work

2.1. Datasets for Interactive Coaching and General Activities

A key related challenge in fitness AI is "situated interaction", where models must proactively provide feedback in real-time. The Qualcomm Exercise Videos Dataset (QEVD) introduced by Panchal et al. is a large-scale benchmark designed for this purpose [6]. Their work, which includes the QEVD-FIT-300K short-clip collection, focuses on the temporal challenge of when to deliver streaming feedback.

While our FIAS project used a custom-curated dataset to analyze specific exercise errors, we utilized the diverse "general activities" (e.g., "grabbing a towel," "drinking from a bottle") from QEVD-FIT-300K to train our model's 'idle' class. This ensures our system can distinguish between deliberate exercises and other real-world actions.

Our contribution remains distinct: where Panchal et al. focus on the interactive dialogue layer, FIAS focuses on ensuring the underlying recognition engine is robust to cross-view failure and interpretable via our XAI-LLM pipeline.

2.2. Advancements in Skeleton-Based Action Recognition

Skeleton-based human action recognition has emerged as a prominent and effective approach within computer vision. By representing human motion as a time-series of keypoint coordinates, this modality offers robustness to variations in background, lighting, and camera viewpoints. A paradigm shift occurred with the introduction of graph-based deep learning models, which are uniquely suited to the non-Euclidean topology of the human skeleton.

The foundational work by Yan et al. [9] introduced the Spatio-Temporal Graph Convolutional Network (ST-GCN), a model that effectively learns both spatial dependencies between joints and their temporal dynamics. This spurred a wave of research focused on enhancing predictive power. Models such as ST-GCN++ [4] represent ongoing efforts, often incorporating more sophisticated network designs to achieve state-of-the-art accuracy on large-scale benchmarks. However, the predominant focus of this research has been on improving classification metrics, with less emphasis on the XAI Interpretability of the models' decision-making processes.

2.3. The Need for XAI Interpretability in Deep Learning Models

While advanced models such as ST-GCN++ [4] have demonstrated impressive performance, their black-box nature poses a major obstacle to adoption in applications that demand transparency and user trust—such as automated fitness coaching. The field of eXplainable AI (XAI) [1] has introduced numerous techniques aimed at revealing the internal mechanisms of deep neural networks. Among these, Gradient-weighted Class Activation Mapping (Grad-CAM) [8] has emerged as a widely used method for visualizing model attention. By producing heatmaps that highlight salient input regions, Grad-CAM offers intuitive visual explanations of model predictions. Although its effectiveness has been well established in image-based tasks, applying it to the spatio-temporal domain of skeleton-based action recognition remains an emerging research direction. Most prior studies have focused primarily on improving model accuracy, leaving the explainability aspect largely underexplored. This paper aims to bridge this gap, emphasizing that for Human Action Recognition (HAR) systems to be truly effective, they must also provide interpretable rationales for their predictions.

Recent research has moved beyond simply applying XAI methods to skeleton-based HAR [7], instead beginning to critically evaluate their reliability and the metrics used for assessment. This is a crucial development, as the black-box nature of models like ST-GCN++ continues to hinder trust in high-stakes applications such as fitness guidance. For example, Pellano et al. examined the applicability of established XAI evaluation metrics—specifically faithfulness and stability—to explanations generated by CAM and Grad-CAM on 3D skeleton data. Their findings reveal the complexity of this challenge: faithfulness can be inconsistent across contexts, whereas stability tends to serve as a more reliable indicator. This highlights the ongoing need for robust evaluation standards for explainability in this domain. While such studies focus on validating XAI metrics, our proposed framework, FIAS, extends this line of work by demonstrating a practical application of explainability—leveraging Grad-CAM outputs to power a downstream LLM-based coaching tool for interpretable and user-trustworthy feedback.

3. Methodology

Our Fitness Insight Analysis System (FIAS) is implemented as a multi-stage pipeline designed to transform a raw video of an exercise into an explainable, human-centric diagnostic report. The overall architecture of this pipeline is illustrated in Figure 1.

The process begins with a raw **Video** input. In the first stage, we employ a high-performance pose estimator,

RTMPose, to extract 2D skeleton **Keypoint data** for each frame. This data is then formatted into an annotation file suitable for our action recognition model.

In the second stage, the formatted keypoint data is fed into a trained **STGCN++ Action Recognition** model, which performs two critical tasks simultaneously. It first infers the most likely **Predicted class** for the action (e.g., 'squat-correct'). Concurrently, we utilize the backpropagated gradient information from this prediction to perform a **GradCAM gradient analysis**. This analysis generates a spatio-temporal saliency map, identifying the key joints and moments that were most influential to the model's decision.

Finally, in the third stage, both the 'Predicted class' and the saliency data from GradCAM are synthesized in a **Structured Prompt generation** module. This prompt, containing all the relevant biomechanical evidence, is then passed to a **Large Language Model (LLM)**, which is tasked with generating the final, human-readable diagnostic report. This end-to-end framework successfully bridges the gap between a quantitative model prediction and a qualitative, expert-level coaching insight.

3.1. Dataset Curation

Recognizing the limitations of existing public datasets, we curated a custom video dataset for analyzing fitness movements. The dataset comprises approximately 1,600 video clips of three foundational actions: lunges, squats, and push-ups. We defined nine distinct classes, categorizing each action into a "correct" form and two common error variants (e.g., `lunge-correct`, `lunge-too-high`, `lunge-knee-pass-toe`).

A central feature is the systematic control of camera viewpoints. Approximately half of the videos were captured from front-facing angles ($0°, \pm 45°$), while the remainder were from back-facing angles ($180°, \pm 135°$). This bifurcated structure is crucial for evaluating a model's ability to generalize across unseen viewpoints.

3.2. Preprocessing

Our preprocessing pipeline consists of three stages:

1. **Skeleton Extraction:** We extracted 2D skeletons using the RTMPose model [5], represented as 17 COCO-compliant keypoints per frame.

2. **Temporal Sampling:** We uniformly sampled 40 frames from each clip to create fixed-length sequences.

3. **Spatial Normalization:** We normalized each skeleton by translating it so the central hip joint is anchored at the origin (0,0).

For experiments requiring 3D data, we employed the MMPose 3D lifter [2] to estimate a z coordinate for each 2D joint, acknowledging that this process can introduce estimation noise.

3.3. Action Recognition Framework

We utilized the MMAction2 toolbox [3] for all training and evaluation.

3.3.1 Core Recognition Model and Modalities

We selected ST-GCN++ [4] as our primary recognition model. A skeleton at frame t is a graph $G_t = (V_t, E)$, where vertices $V_t = \{v_{t,i} \mid i = 1, \ldots, N\}$ are the joint coordinates. For 2D, $v_{t,i} = (x, y) \in \mathbb{R}^2$; for 3D, $v_{t,i} = (x, y, z) \in \mathbb{R}^3$. We derive four input modalities:

- **Joint**: Raw coordinates $(V_1, \ldots, V_T)$.
- **Joint Motion**: Temporal difference $\Delta V_t = V_{t+1} - V_t$.
- **Bone**: Vectors $b_{t,i,j} = v_{t,j} - v_{t,i}$ for connected joints.
- **Bone Motion**: Temporal difference of bone vectors $\Delta B_t = B_{t+1} - B_t$.

3.3.2 Training Details and Augmentations

Models were trained for **20 epochs** using a **batch size of 16**. We utilized an **SGD** optimizer (lr=0.001, momentum=0.9, weight decay=0.0005, nesterov=True) and a **CosineAnnealingLR** scheduler ($T_{\text{max}} = 20$, $\eta_{\text{min}} = 0$). To enhance model robustness against variations in scale and orientation, we also injected new augmentation methods not native to MMAction2, including **Random Rotation** and **Random Scale** on the keypoint data.

3.3.3 Cross-View Evaluation Protocol

To directly measure viewpoint generalization, we defined three data partitions: a `Front` set, a `Back` set, and a `Both` set. We then trained our models under three distinct schemes and evaluated them on all test sets to measure performance on seen and unseen viewpoints, using Top-1 accuracy as the primary metric.

3.4. Interpretable AI for Automated Coaching

To transform our classifier into a practical fitness tool, we developed a novel XAI pipeline [1].

1. **Saliency Analysis:** We apply Grad-CAM to our trained ST-GCN++ model to produce spatio-temporal saliency maps, highlighting which body parts were most influential in the model's decision.

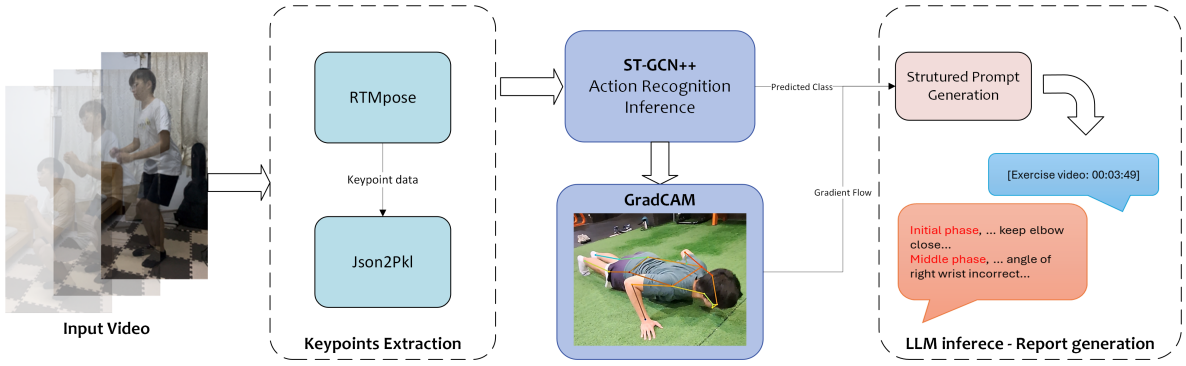


Figure 1. The overall architecture of the FIAS pipeline. From a raw video input, the system extracts keypoints, performs action recognition and Grad-CAM analysis concurrently, generates a structured prompt with the findings, and uses an LLM to produce the final human-readable coaching report.

	Trained on Front			Trained on Back			Trained on Both		
Modality (Dimension)	**Test: Front (f/f)**	**Test: Back (f/ba)**	**Test: Both (f/bo)**	**Test: Back (ba/ba)**	**Test: Front (ba/f)**	**Test: Both (ba/bo)**	**Test: Both (bo/bo)**	**Test: Front (bo/f)**	**Test: Back (bo/b)**
2D Joint	0.9452	0.3250	0.6023	0.9125	0.3288	0.5906	0.9532	1.0000	0.8375
2D Joint Motion	0.9589	0.2750	0.5380	**0.9625**	0.2466	0.5556	**0.9649**	1.0000	0.8500
2D Bone	0.9452	0.3750	0.6842	**0.9625**	0.3699	0.5906	0.9240	1.0000	0.8125
2D Bone Motion	**1.0000**	0.2375	0.6140	**0.9625**	0.3562	0.5848	**0.9649**	1.0000	0.8250
3D Joint	0.7808	0.2500	0.5263	0.8375	0.2192	0.5556	0.8129	0.9178	0.8875
3D Joint Motion	0.7945	0.2875	0.5556	0.7750	0.2055	0.6082	0.8187	0.8904	0.9250
3D Bone	0.8082	0.2250	0.4678	0.8625	0.2603	0.5731	0.7953	0.9315	0.9000
3D Bone Motion	0.7397	0.2875	0.5029	0.8125	0.1918	0.5789	0.7602	0.8493	0.9000

Table 1. Cross-View Generalization Results (Top-1 Accuracy). This table compares the performance of different modalities when trained on front, back, or both views, and tested across all three conditions. The highest accuracy in each primary test column (f/f, ba/ba, bo/bo) is highlighted in bold.

2. **Automated Report Generation:** Our system programmatically analyzes these maps to extract key information (e.g., most activated joints). This structured data is passed to an LLM via an engineered prompt, instructing it to act as a sports science expert and synthesize the data into a concise diagnostic report, avoiding all technical jargon.

4. Experiments and Results

We now present the results of our comprehensive evaluation. Our findings reveal a clear path to achieving robust action recognition by prioritizing training strategy and data representation.

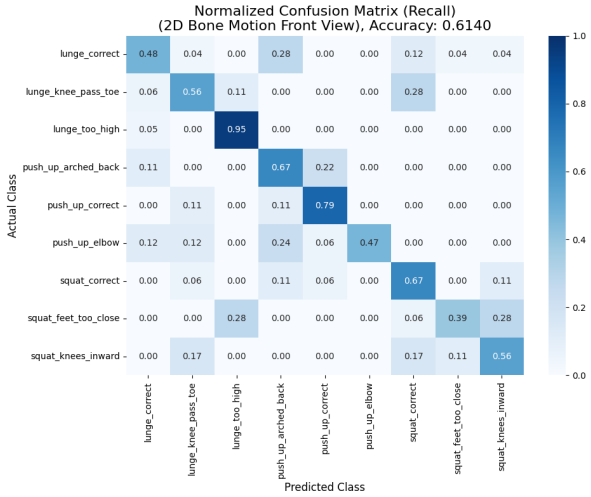
4.1. Cross-View Generalization Analysis

The core of our investigation lies in understanding how models generalize across viewpoints. The results for all modalities and training schemes are detailed in Table 1.

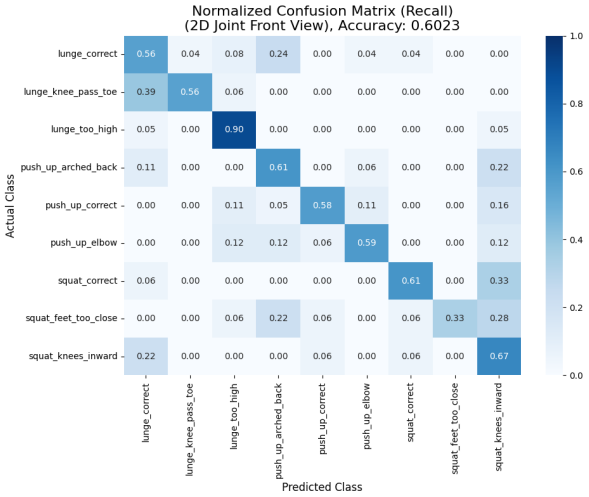
Motion and Bone Features are Highly Effective. Our results indicate that derived 2D features representing dynamics are the most powerful. The top-performing models consistently utilize these modalities, with `2D Bone Motion` achieving a perfect 100% accuracy on the front-view test (f/f), and `2D Joint Motion` and `2D Bone Motion` achieving the highest score of **96.49%** on the mixed-view test (bo/bo).

Single-view models exhibit severe degradation. The data confirms the critical problem of viewpoint overfitting. As shown in Figures 2a and 2b, models trained on a single view exhibit significant error diffusion. Even the perfect-scoring `2D Bone Motion` model (100% on f/f) sees its accuracy plummet to 23.75% when tested on the back view (f/ba). This substantial degradation indicates their unsuitability for real-world applications.

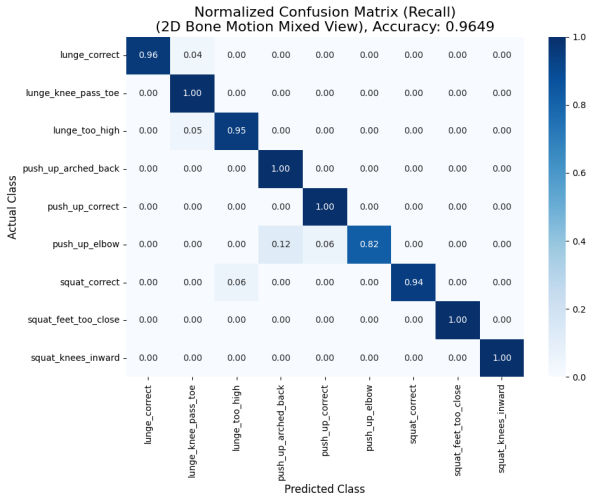
Mixed-View Training Creates Highly Robust Models. Our findings strongly suggest that a mixed-view training strategy provides the most robust generalization. Models trained on the 'Both' dataset not only achieve the highest accuracies but also develop a remarkable ability to generalize. As shown in Table 1, all four 2D models trained on mixed data achieve a perfect 100% accuracy on the front-view test set (bo/f) and very strong accuracies (81-85%) on the back-view test set (bo/b). This demonstrates that the model has learned a truly robust, viewpoint-invariant representation. The confusion matrix for our 'Best Overall 2D Model' in Figure 2c vividly illustrates this, presenting a near-perfectly diagonalized matrix.



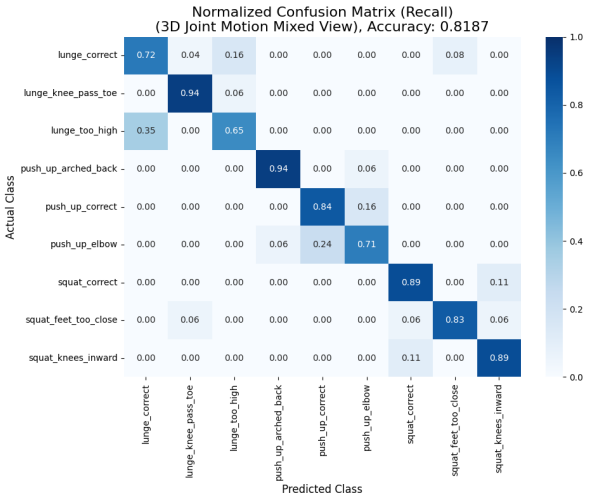
(a) Best Single-View Model (2D Bone Motion, Front View)



(b) Baseline Model (2D Joint, Front View)



(c) Best Overall Model (2D Bone Motion, Mixed View)



(d) Best 3D Model (Joint Motion, Mixed View)

Figure 2. Confusion Matrices for Representative Models. The ‘Baseline‘ and ‘Best Single-View‘ models show significant error diffusion. The mixed-view ‘Best 3D Model‘ improves to a block-diagonal structure, while the ‘Best Overall 2D Model‘ achieves near-perfect diagonalization, indicating superior classification accuracy and fewer Inter-class confusions under XAI analysis.

Modality	2D Model	3D Model	2D Model with Categorical Loss and augmentation	2D Model with augmentation
Joint	0.9532	0.8129	0.9812	0.9812
Joint Motion	0.9649	0.8187	0.9812	0.9953
Bone	0.9240	0.7953	0.9671	0.9765
Bone Motion	0.9649	0.7602	0.9812	0.9859

Table 2. Comparison of Model Accuracy Across Different Input Modalities and Augmentation Strategies. Our augmentations (Random Rotation, Random Scale) were the primary driver of high accuracy. As shown in Table 2, augmentations alone boosted our 2D models to near-perfect Top-1 accuracy (e.g., 99.53% on Joint Motion)

4.2. The Decisive Advantage of 2D Data

A key conclusion from our analysis is that **2D modalities are conclusively superior to 3D** for this task. In every

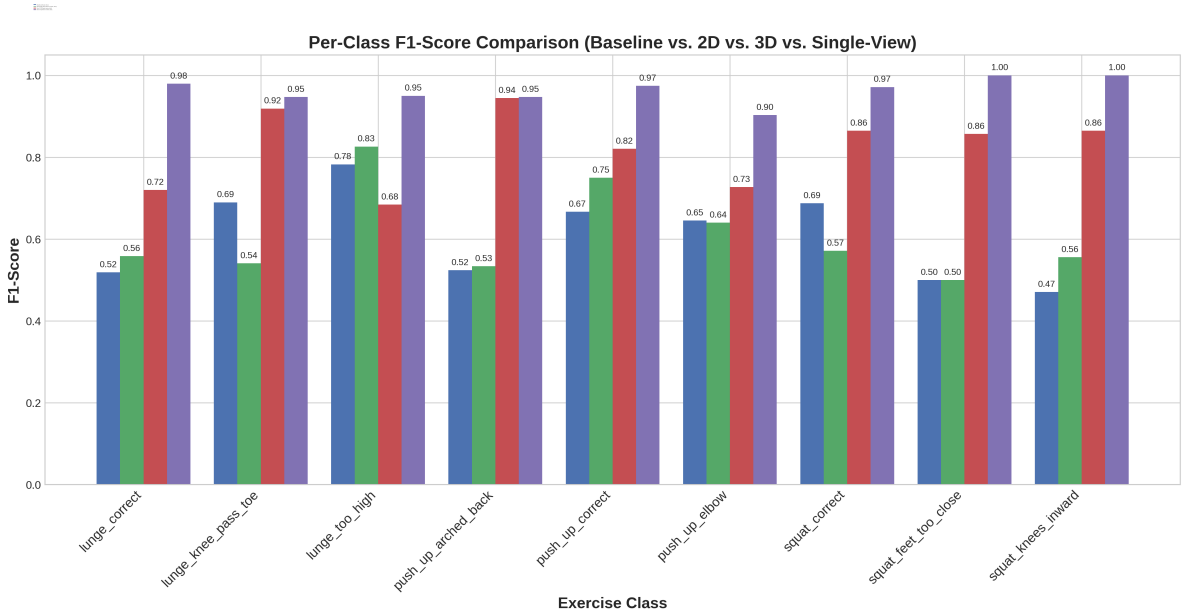


Figure 3. Per-Class F1-Score Comparison. The chart illustrates F1-scores for individual exercise classes across four representative models evaluated on the challenging mixed-view 'both' test set, highlighting the superior performance of the 'Best Overall 2D' model.

Class	Avg. Stability	Avg. Responsiveness	Temporal IoU
`lunge`	85.86%	**0.408s**	0.494
`push_up`	76.68%	0.956s	**0.504**
`squat`	**44.93%**	**2.358s**	0.437
`idle`	N/A	**0.099s** (to idle)	N/A
OVERALL	**64.47%**	**0.670s** (Avg)	**0.478** (mIoU)

Table 3. Analysis of Per-Category Real-Time Performance for the Proposed Action Recognition Model. Model is highly effective at identifying the general action in real-time. Its main challenge is distinguishing between subtle variations of that action.

matched-view scenario ('f/f', 'ba/ba', and 'bo/bo'), the top-performing 2D models achieve significantly higher accuracy than their 3D counterparts. For instance, in the mixed-view test ('bo/bo'), the best 2D model reaches 96.49% while the best 3D model only manages 81.87%. This is likely because 3D pose data inferred from a single 2D camera is not truly viewpoint-invariant and carries view-dependent biases. The simpler 2D representations prove more robust, leading to the conclusion that a superior training methodology with effective 2D features is the most impactful path to success.

4.3. Per-Class Performance and Qualitative Analysis

To gain a more granular understanding, we performed a per-class analysis using the F1-score, visualized in Figure 3. The chart starkly illustrates the inability of single-view models to generalize, as they exhibit low performance across nearly all classes on the mixed-view test. The massive performance leap with mixed-view training demonstrates that data diversity is the most crucial factor for building a robust system.

Superiority of the Optimal 2D Model

The 'Best Overall 2D model' is demonstrably superior to the 'Best 3D model', achieving perfect or near-perfect F1-scores on most classes. The performance gap is particularly evident in classes like 'push-up-correct' (0.97 for 2D vs. 0.75 for 3D), providing conclusive visual proof that a well-trained 2D model decisively outperforms its 3D counterpart.

Anomaly in 'lunge-too-high' Class

An interesting exception was observed in the 'lunge-too-high' class, where the 'Best Single-View 2D' model (F1=0.83) outperformed the 'Best 3D (Mixed-View)' model

(F1=0.68). We hypothesize this is due to a highly discriminative 2D cue (e.g., thigh angle) that is exceptionally clear from the front view, which the specialist model learned as an effective shortcut. The 3D mixed-view model, facing the more complex challenge of creating a unified representation, may have produced a more generalized but less effective solution for this specific case.

Analysis of Real-Time Metrics

At the category level, the model's overall performance is respectable. An overall Stability of 64.47% and a mean Temporal IoU (mIoU) of 0.478 (where $\sim$0.5 is a common threshold for correct detection) prove that the model can successfully identify the correct exercise type (e.g., 'lunge') for the majority of the duration. Furthermore, the overall Responsiveness of 0.670s shows the model is fast at detecting when a general action begins (Tab. 3).

However, the results also highlight a clear bottleneck: the 'squat' category. Its Stability is low (44.93%) and its Responsiveness is slow (2.358s). This suggests that while the model is proficient at identifying lunges and push-ups, it struggles to differentiate 'squat' movements from 'idle' at the beginning of an action. In contrast, at the fine-grained level (evaluating all 9 classes), the overall performance is significantly lower (e.g., 33.64% Stability and 4.223s Responsiveness). This discrepancy proves that the model's main challenge is not in identifying the general action, but in distinguishing between its subtle correct and incorrect variations in real-time.

4.4. XAI Case Study: "Push-up" Assessment

To validate our XAI framework, we conducted a qualitative analysis on the "push-up" exercise. We focused on instances correctly classified as `push-up-elbow`, indicating an incorrect elbow position. Grad-CAM visualizations (Figure 4) revealed a consistent pattern of high activation on the wrists, shoulders, and hips. This demonstrates the model's ability to identify the kinetic chain's areas of stress resulting from the primary error. An incorrect elbow alignment leads to compensatory strain on the wrists and instability in the core, reflected by hip activation.

This structured data—the classification and the salient joints—was fed into our LLM pipeline. The LLM, acting as a fitness coach, successfully translated this analysis into diagnostic advice: *"Based on my analysis, it appears there might be an issue with your elbow positioning. This is causing uneven pressure on your wrists and requiring your shoulders and hips to compensate... I recommend focusing on keeping your elbows directly in line with your shoulders..."*

This case study demonstrates the power of our framework to automate the generation of expert-level, human-centric feedback.

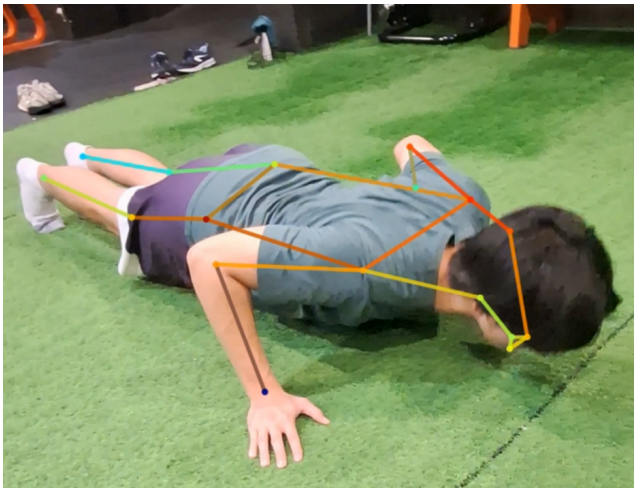


Figure 4. Grad-CAM visualization for the `push-up-elbow` class. The model's attention (warm colors) is focused on the wrists and shoulders, identifying areas of compensatory strain.

5. Conclusion

In this paper, we presented FIAS, a comprehensive framework for building a robust and explainable skeleton-based action recognition system. Our systematic analysis yielded several key insights. We first demonstrated that models trained on single-view data fail catastrophically to generalize, highlighting viewpoint variance as a critical bottleneck.

Our central finding is that prioritizing **2D data diversity** is decisively superior to pursuing more complex 3D-lifted data. We demonstrated that a 2D ST-GCN++ model trained on a mixed-view dataset achieves a state-of-the-art offline accuracy of 96.49%, substantially outperforming 3D-lifted counterparts.

Furthermore, our real-time analysis confirmed the system's practical viability. On an RTX-1080 GPU, the model achieved a strong category-level mean Temporal IoU (mIoU) of 0.478 and a rapid average responsiveness of 0.670s, proving it can generate accurate reports in a streaming context.

Finally, we moved beyond simple classification by introducing a novel XAI pipeline. By utilizing Grad-CAM to analyze the model's prediction activation map, our system gains a temporal and biomechanical understanding of *why* a prediction was made. By integrating this hierarchical understanding with an LLM, FIAS successfully translates complex biomechanical evidence into trustworthy, human-like, and expert-level coaching advice. This work provides a complete blueprint for moving beyond simple classification to create truly intelligent and interpretable

fitness tools.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. 2, 3

[2] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 3

[3] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 3

[4] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition, 2022. 2, 3

[5] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023. 3

[6] Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Bohm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, Mark Todorovich, Ingo Bax, and Roland Memisevic. What to say and when to say it: Live fitness coaching as a testbed for situated interaction, 2024. 2

[7] Kimji N. Pellano, Inga Strümke, and Espen Alexander F. Ihlen. From movements to metrics: Evaluating explainable ai methods in skeleton-based human activity recognition, 2024. 2

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019. 2

[9] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018. 2