# Adapting Explainable AI methods for multi-target tasks

## Addressing challenges and Inter-Keypoint dependencies in Cricket Pose Analysis

**Atanas Semov**
**Supervisors: Ujwal Gadiraju, Danning Zhan**
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Atanas Semov
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Danning Zhan, Mark Neerincx

**Abstract**

Pose estimation models predict multiple interdependent body keypoints, making them a prototypical example of multi-target tasks in machine learning. While existing explainable AI (XAI) techniques have advanced our ability to interpret model outputs in single-target domains, their application to structured outputs remains underdeveloped. This work investigates how XAI methods can be adapted to explain pose estimation models, particularly in the context of cricket shot analysis. Guided by three research questions, we identify key challenges such as capturing inter-keypoint dependencies and providing interpretable explanations of structured outputs. We analyze both geometric and heatmap-level behavior of a pretrained pose estimation model to distinguish between two cricket shots - the pull and the cover drive. Through techniques like cosine similarity on heatmaps and polynomial trajectory modeling, we reveal how the model internally differentiates between similar motion patterns. Our framework introduces novel techniques for inter-keypoint explanation, contributes domain-specific insights into model behavior, and demonstrates the feasibility of interpretable structured predictions in high-dimensional, real-world tasks.

# 1 Introduction

Artificial Intelligence (AI) has made its way into many aspects of modern life - from recommending movies and driving cars to supporting decisions in medicine and finance. Yet, as these systems grow more powerful, they also become more opaque. Understanding why an AI model makes a certain prediction has become just as important as the prediction itself. This is where Explainable AI (XAI) comes in. XAI refers to a set of techniques designed to make the behavior of complex models interpretable to humans. It helps us trust, debug, and improve AI systems, especially in high-stakes settings where decisions have real-world consequences.

So far, most progress in XAI has focused on relatively simple tasks like image classification, where the goal is to assign a single label to an image. But many real-world problems are far more complex. In particular, multi-target tasks, where models make several interdependent predictions at once, pose unique challenges for explanation. One such task is human pose estimation, which predicts the positions of multiple keypoints (like shoulders, elbows, and knees) to understand how a person is positioned in space. This has important applications in areas such as sports analytics, physical rehabilitation, and human-computer interaction.

Existing XAI methods often produce explanations for each output variable independently. However, this approach is insufficient for pose estimation, where the outputs are inherently interconnected. For example, in a cricket action, a movement in the shoulder can affect the trajectory of the elbow, wrist, hips, and even the legs. Understanding these interdependencies is essential for generating explanations that reflect the true reasoning of the model.

In the healthcare domain, researchers have begun to tackle similar challenges. For multi-target medical tasks, such as predicting multiple diagnoses or outcomes from multimodal inputs, XAI techniques for multi-output models and multimodal XAI frameworks have been effectively employed. These methods enable interpretability by jointly analyzing how input features contribute to multiple correlated outputs, offering insight into complex decision-making processes [1, 2].

Parallel developments in computer vision, such as XPose, have introduced group-based attribution techniques like Group Shapley Value (GSV) and Group-based Keypoint Removal

(GKR) to interpret how clusters of keypoints contribute to pose predictions [3]. These approaches represent significant progress, but open questions remain about how such methods can be adapted for domain-specific applications, particularly in sports contexts like cricket, where biomechanical dependencies play a crucial role.

This research addresses the question: **How can XAI methods be adapted for multi-target tasks like pose estimation?** To explore this, the study is guided by three sub-questions:

1. What are the specific challenges in applying XAI methods to multi-target tasks?
2. How can interdependencies between keypoints be identified and explained?
3. How can XAI be used to understand the behavior of the model in relation to the estimation of the cricket pose?

The main contributions of this work include a critical evaluation of existing XAI methods, identification of effective adaptation strategies, and the development of a framework for generating and assessing multi-target explanations. This framework aims to produce explanations that are not only accurate but also informative for both technical and domain-expert audiences.

# 2 Related Work

As AI systems are increasingly used in real-world settings, it's no longer enough for them to just make accurate predictions-we also need to understand how and why they arrive at those decisions. One of the most widely used tools for understanding model behavior is SHAP, which stands for SHapley Additive exPlanations [4]. SHAP is based on ideas from game theory and assigns each input feature a score that shows how much it contributes to the final prediction. However, this technique is not directly applicable to models with structured, multi-target outputs, such as human pose estimation, where outputs (e.g., keypoints) are spatially and semantically dependent.

## 2.1 Gradient-Based XAI for Structured Outputs

In pose estimation, understanding the interdependence between predicted keypoints is crucial for building meaningful explanations. One of the most relevant works in this domain is XPose [3], which extends SHAP to structured outputs by introducing Group Shapley Values (GSV) and Group-based Keypoint Removal (GKR). GSV measures the contribution of groups of joints (e.g., shoulder-elbow-wrist) to the pose prediction, while GKR evaluates the model's sensitivity to their removal. This approach captures spatial dependencies across joints and makes the attribution process more reflective of human anatomy and movement. XPose demonstrates that structured, group-level explanations are more informative than per-keypoint attributions, especially in dynamic activities such as sports or motion analysis.

Building on the idea of spatial attribution, PoseIG [5] applies Integrated Gradients to pose estimation. Instead of grouping keypoints, it focuses on generating individual saliency maps for each joint prediction. These maps highlight the image regions that most influence each keypoint, and the authors introduce evaluation metrics to quantify the strength and consistency of these attributions. It offers a fine-grained view into model reasoning and is particularly helpful for identifying failure modes such as occluded or mislocalized joints.

While SHAP and Integrated Gradients offer general frameworks, visual domains like medical imaging often rely on more spatially intuitive methods such as Grad-CAM [6]. Li et al. [7] employ Grad-CAM in a multi-label retinal disease classification model built from VGG19 and ResNet50. Grad-CAM heatmaps are used not only to interpret predictions post-hoc, but also to guide attention-based augmentation–focusing training on relevant lesion areas. This gradient-based saliency improves both accuracy and interpretability, showing how visual explanations can be integrated into the model pipeline.

A similar use of gradient-based attribution appears in UnboxAI by Zhang et al. [2], which addresses a multimodal, multi-task medical setting. The authors introduce a fusion-aware framework where gradients from image, text, and time-series inputs are aligned with different clinical prediction tasks. Saliency maps are used to explain how each modality contributes to each task, and alignment losses are added to ensure consistency across modalities. This cross-task interpretability ensures that the model remains transparent even in highly complex settings involving interrelated outputs.

### 2.2 Summary and Relevance to This Work

Together, these works demonstrate how XAI methods, especially those leveraging gradient-based saliency, can be adapted to multi-target and structured prediction tasks. XPose shows that dependencies between outputs must be made explicit through group-wise attribution, while PoseIG provides a scalable method to assign spatial importance per keypoint. Li et al. and Zhang et al. highlight the value of gradient-based explanations in improving interpretability and performance in high-dimensional medical tasks.

Our work builds on these approaches by focusing on pairwise relationships between predicted keypoints in pose estimation. Unlike XPose, which uses Shapley values over joint sets, we quantify heatmap similarity across keypoints using cosine similarity. This method provides a data-driven and computationally efficient way to expose spatial dependencies in the model's output space, and offers a novel lens for structured interpretability in tasks such as cricket pose analysis.

## 3 Challenges and Methodological Approach

The goal of this research is to investigate how explainable AI (XAI) methods can be effectively adapted for multi-target tasks such as human pose estimation. Unlike traditional classification problems that output a single prediction, pose estimation models generate structured outputs-typically the spatial coordinates of multiple body keypoints. These keypoints are inherently dependent: the position of one joint (e.g., the wrist) often constrains or predicts the position of others (e.g., the elbow or shoulder). This interdependence is especially critical in domain-specific contexts like cricket shot analysis, where correct body alignment and motion dynamics underpin successful performance.

Applying XAI to pose estimation presents unique methodological challenges. Classical explanation techniques like SHAP and Grad-CAM are typically designed for scalar outputs and do not natively account for the structured nature or interdependencies of pose keypoints. Moreover, the outputs of pose models are often not directly interpretable to non-expert users, limiting the effectiveness of existing saliency or attribution methods in real-world applications such as coaching or rehabilitation. This section outlines the current limitations, proposes a framework for modeling keypoint dependencies, and motivates the need for task-aware XAI in cricket pose analysis.

## 3.1 Limitations of XAI in Pose estimation

Although recent work has made strides in explaining pose models, several critical limitations remain. First, many post-hoc attribution methods, such as Integrated Gradients (as in PoseIG [5]) or Group Shapley Values (as in XPose [3]) rely on gradient information or feature perturbation but do not fully capture the spatial and semantic relationships between keypoints. These methods often treat each output (keypoint) in isolation, making it difficult to understand how combinations of joints influence the model's overall interpretation of a pose.

Second, most saliency-based techniques focus on visual explanations in the input space (e.g., highlighting pixels or regions), but in pose estimation, the outputs themselves (coordinates) are the primary interest. As a result, traditional heatmaps can be unintuitive or insufficient when the goal is to understand model confidence or reasoning in terms of body structure or biomechanics.

Finally, evaluations of interpretability in this domain remain inconsistent. Few studies offer quantitative metrics for explanation quality, and most user-facing applications (e.g., fitness or ergonomics) are validated on small datasets or with limited user studies. This makes it difficult to assess the reliability, robustness, or domain-transferability of XAI methods in structured output tasks like pose estimation.

This analysis directly addresses our **first research subquestion**: *What are the specific challenges in applying XAI methods to multi-target tasks?* The limitations outlined in the literature review - particularly the lack of inter-keypoint dependency modeling, the mismatch between input-focused saliency and structured output interpretation, and the absence of standard evaluation practices - highlight why existing approaches are insufficient for pose estimation. These challenges motivate the development of more targeted interpretability tools that explicitly account for the structured, interrelated nature of pose model outputs.

## 3.2 Identifying Keypoint Dependencies

A central challenge in explaining pose estimation models lies in understanding how predicted keypoints relate to one another. Since body joints are spatially and biomechanically connected, treating each output independently, as many existing XAI methods do, can obscure important structural dependencies. To address this, we introduce a keypoint-level analysis aimed at quantifying interdependencies between outputs in a pose estimation model.

To investigate these relationships, we make use of the heatmap outputs generated by a pre-trained pose estimation model. For each input image, the model produces a set of heatmaps, one per keypoint, indicating the predicted spatial probability distribution for that joint. We then analyze the pairwise overlap between these heatmaps across multiple samples to assess how strongly the activation for one keypoint correlates with others. This yields a Keypoint Dependency Heatmap, a symmetric matrix that quantifies how much each keypoint depends on others in the model's internal representations.

This method allows us to move beyond per-keypoint saliency and instead build a structured understanding of the model's internal pose logic. Keypoints with high mutual dependency values likely reflect strong spatial or functional ties, such as those between the shoulder and elbow, or the knee and hip. Conversely, weak dependencies may indicate either true anatomical independence or model uncertainties.

We chose this methodology because standard attribution techniques are designed for scalar outputs or categorical decisions and do not capture inter-output relationships. Heatmaps, however, offer a spatial probability view of each joint's prediction, and analyzing their overlaps provides a more natural and interpretable way to uncover how the model encodes structural dependencies. This approach aligns better with the underlying nature of pose estimation and bridges the gap left by traditional explanation tools.

This analysis directly supports the **second research subquestion**: *how can interdependencies between keypoints be identified and explained?* By visualizing and quantifying these internal relationships, we lay the foundation for building more structured, context-aware explanation mechanisms that reflect the true nature of human motion.

### 3.3 Understanding Pose Estimation in Cricket Shot Analysis

To evaluate how explainable AI methods can support domain-specific interpretation, we apply our keypoint-based analysis to a focused case study: distinguishing between two fundamental cricket shots - the pull shot and the cover drive. These actions involve different biomechanical patterns, particularly in the movement of the arms, which makes them a suitable testbed for analyzing pose estimation behavior and its interpretability.

This approach begins by examining the dynamic relationship between the wrist and the shoulder during each shot type. Using the keypoint coordinates generated by the pose estimation model, we track the relative spatial position of the wrist with respect to the shoulder over the sequence of frames. This relationship is formalized through mathematical expressions, which quantify angles or displacements in 2D space. By characterizing these movement profiles, we establish pose-based descriptors that differentiate the two shot types based on arm motion alone.

In addition to geometric analysis, we also investigate the internal behavior of the pose model by comparing the predicted heatmaps for the same keypoints across the two shot categories. This allows us to observe how confident the model is in predicting certain joint positions, and whether its attention shifts across different joints depending on the shot. For instance, we observe whether the model assigns higher spatial probability to the front wrist in a pull shot compared to a cover drive, or whether certain joints become less localized in one context versus another.

This methodology was chosen because it aligns closely with the challenges identified in explaining structured pose outputs. Rather than treating the classification of sports actions as a black-box outcome, we break down the model's internal representations into interpretable components - keypoint movement patterns and spatial attention differences. These insights are not easily accessible through traditional XAI methods, which lack the resolution or task-awareness to interpret such fine-grained motor behavior. Our approach offers a domain-relevant pathway for understanding what the model focuses on when distinguishing between complex physical movements.

By combining coordinate-based motion analysis with heatmap-level interpretation, we gain insight into how the model internally differentiates between similar but functionally distinct poses. This analysis reveals not only what the model predicts, but how it arrives at those predictions, thereby advancing the broader goal of making structured output models more transparent in task-specific applications like sports pose assessment.

## 4 Analysis of Keypoint Behavior Across Cricket Shots

This section presents an empirical analysis of how a pose estimation model distinguishes between two common cricket shots: the pull shot and the cover drive. Building upon the methods outlined in the previous section, we investigate how different keypoints contribute to the model's internal decision-making and whether the spatial configurations and heatmap activations reveal consistent pose-level differences. Our goal is to interpret the model's behavior by focusing on visual and geometric evidence derived from its outputs.

### 4.1 Keypoint Activation Contrast Between Shots

The first axis of analysis focuses on the predicted heatmaps for individual keypoints. By comparing these heatmaps across the two shot types, we assess how the model allocates spatial attention and confidence across joints. Each heatmap represents the model's spatial probability distribution for a given joint, allowing us to visually inspect where the model "looks" when predicting joint positions.

We observe that the model exhibits stronger and more focused activations in specific joints depending on the shot type. For example, in pull shots, the wrist of the dominant hand often displays a sharper and more localized heatmap, likely due to its extended lateral movement across the horizontal axis. In contrast, during cover drives, shoulder and elbow activations tend to be more prominent and stable, reflecting the controlled vertical extension of the leading arm.

These differences suggest that the model adapts its internal representation depending on the type of motion, emphasizing keypoints that are most informative for each shot. This finding supports the idea that pose estimators implicitly encode action-level information even when not explicitly trained for classification, and that this information can be revealed through targeted heatmap analysis.

### 4.2 Geometric Signatures of Arm Movement

In parallel with heatmap analysis, we also explore the geometric patterns of arm motion by analyzing the spatial relationship between the wrist and the shoulder. For each shot type, we compute the relative wrist position with respect to the shoulder over a sequence of frames, using 2D displacement vectors and angle calculations to formalize these movements.

Our analysis shows clear distinctions between the two shots: pull shots typically involve a wide horizontal sweep of the wrist, often with a lower vertical displacement, while cover drives exhibit a more diagonal or vertical movement path, consistent with the upward and outward drive of the bat. These patterns are captured through trajectory plots and angle measurements, which consistently reveal higher angular elevation in cover drives and wider horizontal spans in pull shots.

This geometric analysis complements the heatmap-based interpretation by showing how the model's predictions correspond to biomechanically meaningful differences in joint movement. By capturing these patterns through interpretable mathematical descriptors, we not only understand what the model predicts but gain insight into how these predictions align with real-world motion characteristics. This dual perspective offers a richer view of model behavior in a sports-specific pose estimation context.

## 5 Experimental Setup and Results

Having outlined the methodological approach and the motivations behind it, we now turn to the practical evaluation of these ideas. This section presents the experimental setup and results obtained through our analysis of pose estimation in multi-target tasks. We describe the datasets, preprocessing steps, and models used in our experiments, followed by a detailed presentation of the key findings. The goal is to assess how well the proposed interpretability techniques reveal structured dependencies between keypoints and explain the model's decision-making process, particularly in domain-specific contexts such as cricket shot recognition.

## 5.1 Experimental Setup

### Datasets

To evaluate the interpretability of pose estimation in multi-target tasks, we conduct experiments in two settings: general human movement analysis using the JHMDB dataset and domain-specific sports analysis using the CKT Cricket dataset [8, 9].

The JHMDB dataset consists of short human action video clips and is commonly used in pose and action recognition research. In our experiments, we use only the raw video content and do not rely on the original annotations. Our objective is not to evaluate pose prediction accuracy against ground-truth labels, but rather to study the internal behavior of a pretrained pose estimation model through keypoint heatmaps and interdependencies.

For the cricket-specific experiments, we use the CKT Cricket dataset, a publicly available collection of cricket video clips sourced from broadcast footage. This dataset does not contain pose or action annotations.

### Preprocessing

To isolate the batter and reduce background noise, we apply a preprocessing step using YOLOv8 [10], an efficient object detection model released by Ultralytics, to generate bounding boxes around the batter as seen in Fig. 1. YOLOv8 uses an anchor-free detection head and advanced backbone architecture, achieving an excellent balance of accuracy and real-time inference (>50 FPS), making it ideal for cropping batter frames from cricket videos efficiently and accurately.

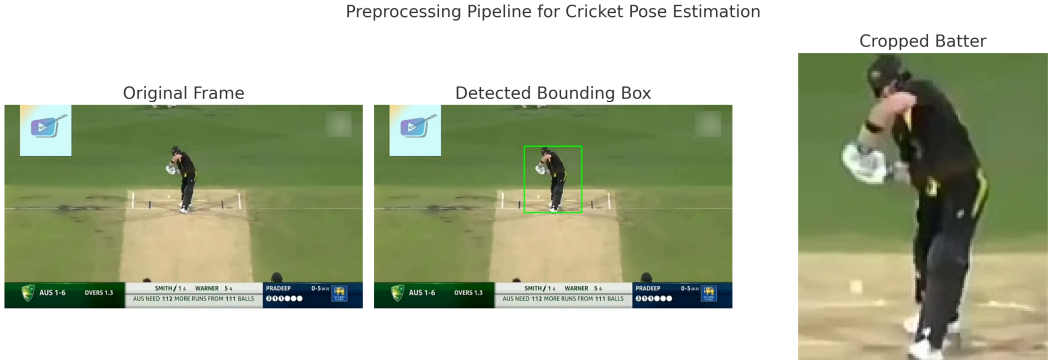

Figure 1: *Cricket shot preprocessing stages.* Left: original input frame. Center: YOLOv8 detection with bounding box. Right: cropped output used for pose estimation.

Since the batter's position may shift across frames-due to camera movement or zoom-we apply this detection and cropping process at regular frame intervals to ensure consistent localization. Additionally, we manually trim the beginning and end of each cricket video clip to exclude irrelevant frames and retain only the core action segment. This step ensures that the pose estimation model focuses exclusively on meaningful motion patterns associated with the cricket shot.

### Model

For pose estimation, we use the pretrained `keypointrcnn_resnet50_fpn` model from PyTorch's torchvision library [11]. This model is built on top of Faster R-CNN with a ResNet-50 backbone and a Feature Pyramid Network (FPN), and it is trained on the COCO keypoint dataset. It provides robust multi-scale feature extraction and outputs 17 keypoints per person, each with an associated heatmap. This architecture is well-suited for human motion analysis and has demonstrated high accuracy on diverse pose datasets.

In both experimental settings, we aggregate predictions across multiple frames and video samples to improve result reliability and minimize the effect of temporary occlusions or outlier frames. For example, when computing keypoint dependencies or visualizing average heatmap patterns for a specific shot type, we average outputs across several sequences. This technique produces smoother and more representative insight into the model's internal behavior.

This modular pipeline, combining person detection, pose estimation, and post hoc interpretability analysis, is reproducible and adaptable to other structured movement domains such as sports performance, rehabilitation, or biomechanics.

## 5.2 Literature Context: XAI in multi-target tasks

The first research subquestion of this work - *What are the specific challenges in applying XAI methods to multi-target tasks?* - is grounded in the complexity of explaining models whose outputs are not only numerous but often highly interdependent. Across domains like healthcare and computer vision, traditional XAI techniques such as SHAP and Grad-CAM have shown utility, yet encounter major limitations when applied to structured or multi-output settings. That is why we should dive deeper into some innovative methods proposed in these fields to explain the output of models.

### Healthcare Domain

In healthcare, for instance, multi-target models frequently predict related diagnoses or biomarkers. To explain them, recent research has extended SHAP to better accommodate multi-target settings. For example, in the survey by Tjoa and Guan [1], several adaptations of SHAP are reviewed, including hierarchical SHAP, which distributes contributions across structured layers (e.g., organ systems or diagnostic categories), and clustered SHAP, which aggregates attributions over groups of outputs to highlight systemic effects. Similarly, Bhattarai et al. [12] applied Deep-SHAP to map relationships between neuroimaging biomarkers and cognitive scores in Alzheimer's disease, offering insight into which brain regions contribute to multiple cognitive targets.

Visual explanation methods like Grad-CAM offer an alternative by highlighting salient image regions responsible for predictions. Grad-CAM operates by backpropagating gradients from a target output through the final convolutional layers of a neural network, generating heatmaps that localize the spatial regions most influential to a given prediction. Unlike SHAP, which quantifies feature importance in an abstract input space (e.g., pixel intensity or lab value), Grad-CAM highlights where in an image the model is "looking" when making decisions, offering more interpretable insights for visual domains. A recent example by Li et al. [7] demonstrated the usefulness of Grad-CAM for multi-label retinal disease detection, integrating it into the model pipeline for interpretability-aware training.

Other works reveal further challenges. Kim et al. [13] and Jin et al. [14] show that while saliency maps can be extended to multi-label or multimodal tasks, doing so often requires architectural changes or additional annotations. Moreover, explanations in these settings tend to prioritize individual targets without contextualizing how decisions for one output may affect or depend on others.

Multimodal explainability efforts such as Zhang et al. [2] underscore this issue: even when using gradient-based attention to create cross-modal saliency maps, the complexity of integrating interpretability across different tasks and data types remains a barrier. Similarly, works like Badr'e and Pan [15] and Shi [16] attempt to interpret shared latent features across outputs, yet still face difficulties in surfacing transparent, interpretable relationships between them.

### Computer Vision

In pose estimation, where each output (keypoint) forms part of a coherent spatial structure, methods like XPose [3] and PoseIG [5] have attempted to address this by introducing group-based Shapley attribution or applying Integrated Gradients at the output level. Complementing XPose, the Pose Tutor system [17] targets the human-in-the-loop aspect of pose correction. It combines pose prediction with visual and linguistic explanations to guide users in correcting their posture. Unlike XPose, which focuses on model-centric attribution, Pose Tutor emphasizes user-facing interpretability. It highlights which keypoints deviate from expected configurations and offers actionable guidance for correction, making it particularly suited for real-time sports training and rehabilitation settings.

Beyond these approaches, several recent studies have further explored model-centric interpretability. TransPose [18] integrates interpretability into the model architecture itself using a Transformer, where attention weights reveal which image patches contribute to each keypoint prediction. This method provide valuable diagnostic tools and help uncover failure modes such as shortcut learning or keypoint misclassification.

Building on user-facing interpretability, CARE [19] formulates pose correction as a counterfactual reasoning problem, prescribing minimal joint angle adjustments required to convert an incorrect pose into a correct one. Similarly, Dibenedetto et al. [20] present a lightweight office posture correction system using post-hoc feature importance to guide ergonomic feedback. These works emphasize actionable, real-time pose improvement and extend explainability into practical domains beyond sports.

In summary, current XAI methods struggle with the following challenges in multi-output settings: (1) capturing interdependencies between outputs; (2) offering human-interpretable representations of these dependencies; (3) scaling explanations across tasks or modalities; and (4) providing consistent, robust interpretability without sacrificing model accuracy or usability. These challenges justify the need for novel, domain-specific approaches, like those proposed in this project, which aim to model and visualize inter-keypoint dependencies to provide deeper insight into structured pose estimation models.

# 6 Results

## 6.1 Quantifying Keypoint Dependencies with Cosine Similarity

To understand how the model internally relates different body keypoints, we compute pairwise similarities between the predicted heatmaps produced by the pose estimation model. Each heatmap reflects the spatial probability distribution of a keypoint's location, as shown in Figure 2 for the right foot. A good similarity measure should capture how similarly two joints are spatially activated, rather than relying on raw intensity.

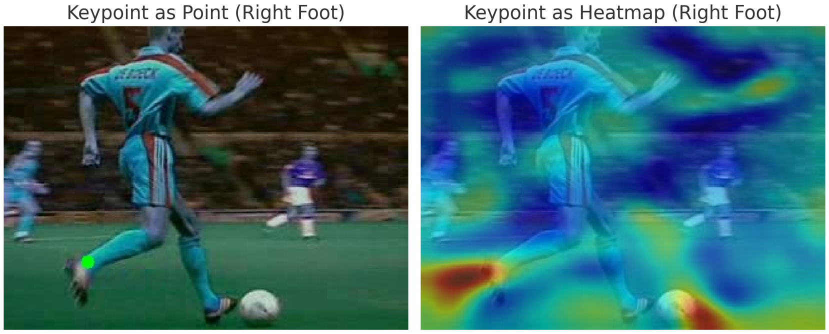
Keypoint as Point (Right Foot)
Keypoint as Heatmap (Right Foot)

Figure 2: *Visualizing Keypoint Representations.* Left: Right foot keypoint shown as a detected point. Right: Corresponding heatmap visualization highlighting spatial confidence. This comparison illustrates how the model represents keypoints not just as coordinates but as spatial distributions over the input image.

We use *cosine similarity*, a widely adopted metric in high-dimensional vector analysis, particularly effective for comparing spatial patterns. It measures the orientation similarity between two vectors, making it robust to differences in magnitude that commonly occur in heatmaps across frames or joints. Unlike metrics such as Euclidean distance or Pearson correlation, cosine similarity is scale-invariant, allowing us to focus on the spatial co-activation patterns.

Given two normalized and flattened heatmaps $\mathbf{A}$ and $\mathbf{B}$, the cosine similarity is defined as:

$$\text{cos\_sim}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

where $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the two vectors, and $\|\mathbf{A}\|$, $\|\mathbf{B}\|$ denote their Euclidean norms.

This value lies in the range $[0, 1]$ for non-negative heatmaps, with higher values indicating greater spatial alignment. By computing this similarity for every pair of keypoints across multiple frames and multiple videos, we construct a *Keypoint Dependency Matrix*, which encodes the degree of interdependence between joints.

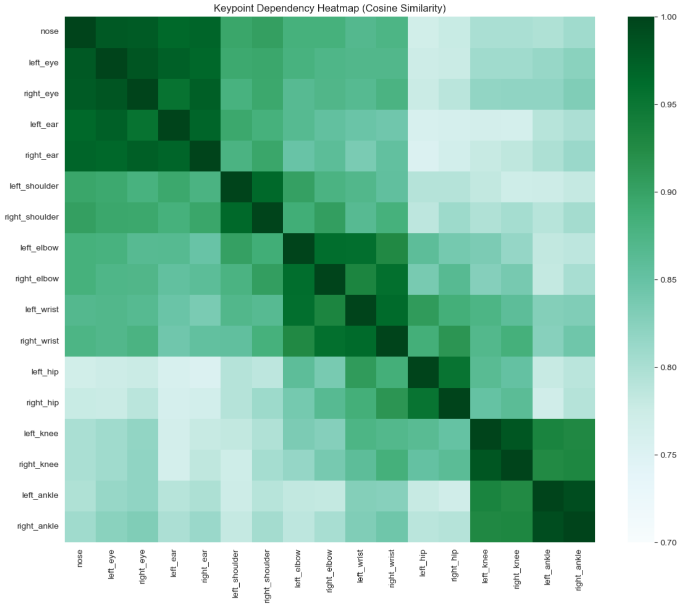

Figure 3: *Keypoint Dependency Matrix based on Cosine Similarity.* Each cell quantifies the average spatial co-activation between two keypoints. Darker regions indicate stronger similarity between the respective heatmaps, suggesting a higher degree of model-level dependency.

As seen in Figure 3, certain keypoints, such as the shoulder and elbow, exhibit strong mutual dependencies, reflecting expected biomechanical relationships. This analysis provides a foundation for understanding the internal logic of pose estimation models and supports structured interpretation of their outputs in both general and task-specific contexts.

## 6.2 Analysis of Keypoint Behavior Across Cricket Shots

This part presents the results of our domain-specific analysis of pose estimation behavior in cricket. Building on the methodology described previously, we examine how the model distinguishes between two common batting techniques: the *pull shot* and the *cover drive*. The analysis is based on two modalities - predicted keypoint heatmaps and coordinate-based movement signatures - with a focus on the left wrist and shoulder as key indicators of arm motion.

To ensure consistency, all videos in each shot category were trimmed to include only the active portion of the movement, thereby removing irrelevant frames. Keypoint predictions and heatmaps were then aggregated per shot type to support comparative analysis across multiple video samples.

### Geometric Analysis of Batter's Form

To establish an interpretable baseline for shot comparison, we begin with a purely geometric analysis of wrist movement relative to the shoulder. This form of analysis mirrors how a human coach or observer would examine a cricket player's motion by visually tracking the arm's path and shape. Importantly, this offers a human-interpretable counterpart to the model-based interpretation methods introduced later.

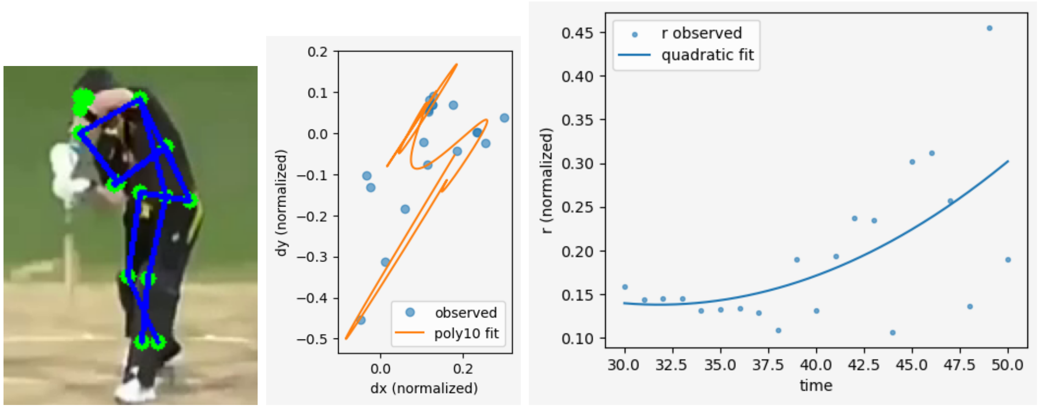

Figure 4: *Comparison of wrist movement during a cover drive.* Left: Pose keypoints detected by the model. Center: Fitted wrist trajectory relative to the shoulder, modeled using a degree-10 polynomial in 2D space. Right: Radial distance between wrist and shoulder over time, fitted with a quadratic curve. The motion exhibits a sweeping horizontal arc and increasing displacement, characteristic of a cover drive.

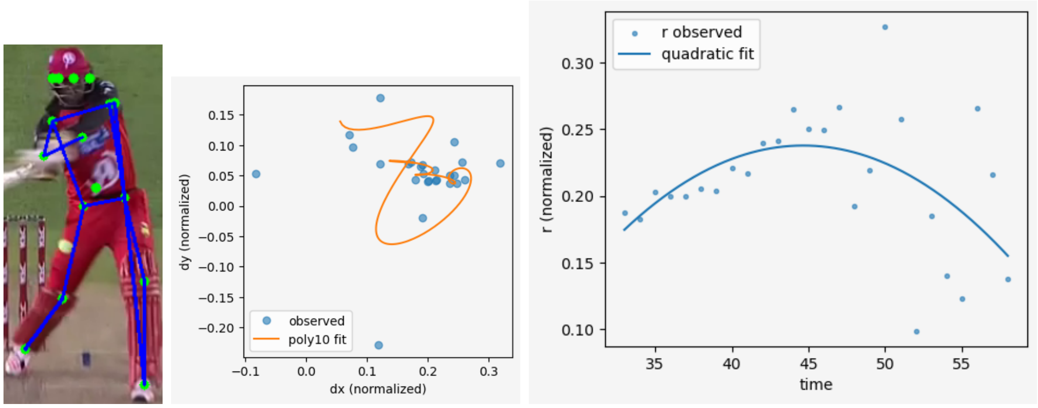

Figure 5: *Comparison of wrist movement during a pull shot.* Left: Pose keypoints detected by the model. Center: Fitted wrist trajectory relative to the shoulder, modeled using a degree-10 polynomial in 2D space. Right: Radial distance between wrist and shoulder over time, fitted with a quadratic curve. The motion exhibits a sweeping horizontal arc and increasing displacement, characteristic of a pull shot.

Each video frame is processed using a pose estimation model, which outputs 17 body joint keypoints. For our analysis, we isolate the left wrist $(x_w, y_w)$ and left shoulder $(x_s, y_s)$ in normalized image coordinates $[0, 1]$. To eliminate the effects of global movement and camera panning, we compute relative wrist coordinates:

$$\Delta x(t) = x_w(t) - x_s(t), \quad \Delta y(t) = y_w(t) - y_s(t)$$

This centers all motion on the shoulder, allowing us to study only the articulation of the arm during the shot.

Then we fit a continuous 2D curve to the wrist trajectory relative to the shoulder using a polynomial model:

$$\Delta x(t) \approx \sum_{i=0}^{d} a_i t^i, \quad \Delta y(t) \approx \sum_{i=0}^{d} b_i t^i$$

Polynomials were selected for their smoothness, differentiability, and analytical tractability. Unlike splines or piecewise models, polynomials offer global control of curve shape, which is especially useful in motion capture where acceleration and velocity trends are of interest.

The choice of polynomial degree $d = 10$ was empirically determined. Lower-degree polynomials ($d \leq 4$) underfit, failing to capture complex wrist behavior such as the sweeping arcs in cover drives or sharp inflections in pull shots. Conversely, degrees beyond 10 tended to overfit and amplify high-frequency noise. Degree 10 provided the best trade-off between expressiveness and robustness, capturing up to two or three distinct changes in motion direction, as illustrated in figures 4 and 5.

To complement the 2D spatial analysis, we further evaluate the radial distance of the wrist from the shoulder as a function of time:

$$r(t) = \sqrt{\Delta x(t)^2 + \Delta y(t)^2}$$

This scalar quantity captures the extension or contraction of the arm during the stroke. A quadratic function was used to model $r(t)$, as this shape generally reflects the biomechanics of a controlled limb motion - an initial rise (extension), peak, and fall (retraction). This fits the cricket context well, where the wrist typically moves outward with the bat and then returns to a neutral position post-contact.

This geometric modeling approach serves multiple purposes:

- It provides a **human-understandable abstraction** of complex limb motion.
- It serves as a **reference point** against which the model's internal attention or uncertainty can be evaluated.
- It reveals clear **shot-specific movement patterns**, suggesting potential for weakly-supervised or unsupervised classification based on motion alone.

By embedding these fitted descriptors into a structured analysis pipeline, we create a bridge between raw model outputs and interpretable biomechanical behaviors. These insights lay the foundation for deeper explanation analysis using heatmaps and attribution metrics in the subsequent sections.

### Model Attention Patterns of Batter's form

To investigate how the pose estimation model internally attends to different keypoints across distinct shot types, we perform a heatmap-level analysis centered on the relationship between the left wrist and the left shoulder. Rather than relying solely on the predicted coordinates, this analysis examines the raw joint heatmaps - spatial probability distributions generated by the model for each keypoint.

For each frame within a temporally aligned clip, we compute the cosine similarity between the flattened heatmaps of the left wrist and the left shoulder. This metric serves as a proxy for joint-level correlation, revealing whether the model's attention toward one joint spatially overlaps or aligns with the other. We track these correlation scores across the time dimension of both shot types as seen in Fig. 6. Given that the initial stance is visually similar in each case, our analysis focuses on how the model's internal representation begins to diverge as the shot motion initiates, revealing whether it can distinguish between the two actions based on movement dynamics.

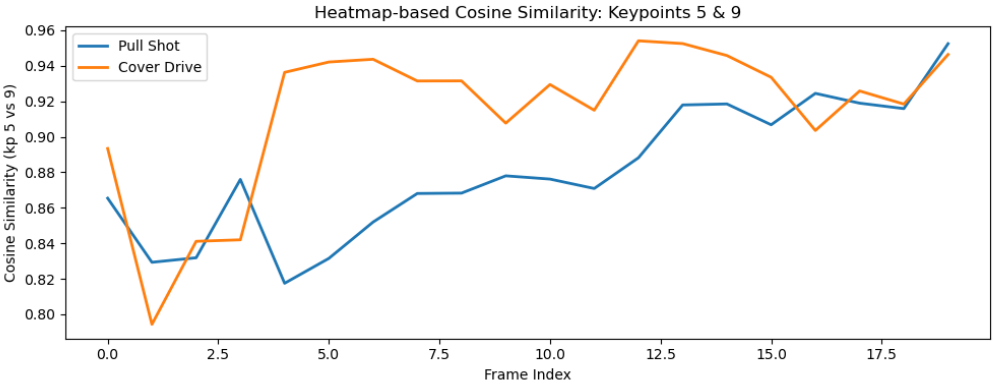

Figure 6: Cosine similarity of predicted heatmaps for keypoints 5 (left wrist) and 9 (left shoulder) across frames. The similarity reflects how similarly the model attends to both keypoints over time for the pull shot and cover drive.

This heatmap correlation analysis bridges the gap between visual joint tracking and latent model behavior, offering a task-relevant lens into how structured pose estimators distribute spatial focus across related outputs in dynamic settings.

# 7 Responsible Research

This project aims to improve the interpretability of pose estimation models in multi-target tasks by analyzing the relationships between predicted keypoints and their behavior in domain-specific contexts, such as cricket shot analysis. In pursuing this objective, we have paid careful attention to issues of ethical responsibility, data transparency, and reproducibility.

## 7.1 Ethical Considerations

All data used in this study comes from publicly available video datasets: the JHMDB dataset for general human movement [8], and the CKT Cricket dataset [9] for sports-specific analysis. These datasets consist of publicly broadcast videos and contain no personal or sensitive information, ensuring compliance with ethical data use standards.

Additionally, we avoid training new models from scratch or on private human subject data, thereby minimizing computational costs and the associated environmental impact of deep learning research. The project focuses solely on the interpretability of existing models rather than expanding predictive capabilities, ensuring that no new biases are introduced in model behavior.

### 7.2 Reproducibility and Model Transparency

All methods used in this research are built on openly available tools and pretrained models. The pose estimation component is powered by the Keypoint R-CNN model with a ResNet-50-FPN backbone [21], available through PyTorch's torchvision library. Similarly, YOLOv8 [10], used for cropping the batter from cricket videos, is an open-source object detector known for its high accuracy and speed.

Despite relying on these public tools, pose estimation pipelines often behave as black boxes due to their complex multi-stage architectures. While popular explanation techniques like Grad-CAM [6] are effective in many image-based tasks, adapting them to pose estimation is challenging. Many pretrained pose models are not designed to expose intermediate gradients in a manner compatible with Grad-CAM, and retraining a custom architecture to facilitate such access is infeasible in this context due to limited data, annotations, and computational resources.

As an alternative, this project focuses on interpreting model outputs directly, specifically, the heatmaps associated with keypoint predictions. These heatmaps provide a natural and meaningful representation of the model's spatial belief for each joint. By quantifying spatial dependencies through cosine similarity across these heatmaps, we provide an interpretable view of how the model perceives structural relationships in human pose, without needing model modifications or retraining.

### 7.3 Code Availability

To support transparency and reproducibility, all code used in this study, including data preprocessing, keypoint extraction, heatmap generation, and dependency analysis, will be made publicly available in a dedicated repository. This will allow other researchers to replicate the experiments, verify results, and adapt the methodology to related domains such as rehabilitation, fitness tracking, or sports analytics.

## 8 Discussion

The frame-wise cosine similarity between the predicted heatmaps of the left wrist and shoulder, as visualized in Figure 6, exhibits patterns that align closely with the radial trajectory profiles observed earlier through polynomial and quadratic fitting. For both shot types, the cosine similarity increases as the shot motion progresses, mirroring the parabolic or rising trend in the radial distance from the wrist to the shoulder over time. This alignment suggests that as the batter transitions from stance to execution, not only does the geometric separation between joints become more pronounced, but the model's focus on those joints also becomes more coordinated, reflecting stronger functional coupling during dynamic motion.

This relationship between geometric displacement and model-based heatmap correlation indicates that the pose estimator internalizes meaningful kinematic dependencies between joints as they become more active in the context of the shot. The increasing similarity can be interpreted as the model reinforcing the biomechanical link between wrist and shoulder as the arm extends and rotates. Interestingly, the correlation peak typically coincides with the frames where the radial movement is at its maximum or inflection, suggesting that the model may rely on such moments to anchor its prediction confidence.

Beyond confirming our earlier trajectory-based findings, this analysis offers further interpretability: it demonstrates that internal saliency patterns in deep pose models reflect not only the spatial configuration of joints but also their temporal coordination. This opens new possibilities for interpreting model behavior in time-sensitive domains such as sports or rehabilitation, where dynamic joint interaction is critical. Moreover, it highlights the potential for using correlation tracking as an explainability signal to identify critical movement phases or diagnose failure modes in model predictions.

# 9 Conclusions and Future Work

This research addressed the overarching question: *How can explainable AI (XAI) methods be adapted for multi-target tasks like pose estimation?* In doing so, we explored three subquestions focused on (1) the specific challenges in applying XAI to multi-target tasks, (2) identifying and explaining inter-keypoint dependencies, and (3) understanding model behavior in structured pose tasks like cricket action analysis.

Our findings revealed several critical limitations of traditional XAI methods when used in structured output settings. Specifically, we showed that approaches like SHAP, Grad-CAM, and Integrated Gradients often fail to capture the spatial and functional dependencies among predicted keypoints. To address this, we proposed and implemented a heatmap-based dependency analysis using cosine similarity, offering a more natural and spatially aware measure of joint correlations.

In addition, we introduced a geometric analysis technique based on polynomial modeling of wrist-shoulder trajectories, enabling us to distinguish between cricket shots based on joint motion patterns. This dual approach - combining coordinate-level motion analysis with heatmap similarity - yielded interpretable and task-relevant insights into model behavior.

**Key contributions include:**

- A novel heatmap-based dependency matrix revealing inter-keypoint relationships.
- A comparative framework for analyzing different action classes using both visual (heatmap) and geometric (trajectory) evidence.
- Domain-specific analysis demonstrating how pose models attend to different joints depending on task context.

**Future work** can extend this study in several impactful directions. First, while our analysis focused on the relationship between the left wrist and left shoulder, examining dependencies among other keypoints, such as between the hips and knees, or across contralateral limbs, could reveal additional biomechanical patterns and further generalize the framework to a wider range of movements and sports contexts. Also, integrating temporal modeling into the interpretability process would allow for the analysis of evolving joint dependencies over the course of an action, potentially uncovering transition dynamics that static analysis cannot capture.

In conclusion, this research demonstrates that adapting XAI to multi-target structured prediction is not only feasible but necessary for understanding complex models in real-world tasks. By aligning explanations with domain-specific patterns and interdependencies, we bring model transparency closer to human reasoning.

# References

[1] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (xai): Toward medical xai,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, Nov 2021.

[2] W. Zhang, D. Wang, and Q. Hu, “Unbox the black-box for medical explainable ai via multi-modal and multi-task learning,” *Information Fusion*, vol. 77, pp. 29–40, Aug 2022.

[3] L. Qiu, Z. Shen, X. Shen, Y. Ge, S. Lu, Z. Li, Y.-W. Tai, C.-K. Tang, G. Yu, and Y. Liu, “Xpose: explainable human pose estimation,” in *ArXiv Workshop*, 2024, available at: arXiv:2403.12370.

[4] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[5] Q. He, L. Yang, K. Gu, Q. Lin, and A. Yao, “Analyzing and diagnosing pose estimation with attributions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 4821–4830.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[7] Z. Li, M. Xu, X. Yang, Y. Han, and J. Wang, “A multi-label detection deep learning model with attention-guided image enhancement for retinal images,” *Micromachines*, vol. 14, no. 2, p. 371, 2023.

[8] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, “A benchmark for human activity understanding in videos,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 459–466.

[9] R. Sharma and M. Sharma, “Ckt cricket pose dataset,” https://www.kaggle.com/datasets/rsrishav/cricket-shot-dataset, 2022, accessed: 2024-06-20.

[10] Ultralytics, “Yolov8: Real-time object detection,” https://docs.ultralytics.com/, 2023, accessed: 2024-06-20.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[12] P. Bhattarai, D. S. Thakuri, Y. Nie, and G. B. Chand, “Explainable AI-based Deep-SHAP for mapping the multivariate relationships between regional neuroimaging biomarkers and cognition,” *European Journal of Radiology*, vol. 174, p. 111403, 2024.

[13] Y. Kim, J. M. Kim, J. Jeong, C. Schmid, Z. Akata, and J. Lee, “Bridging the gap between model explanations in partially annotated multi-label classification,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3408–3417.

[14] W. Jin, X. Li, and G. Hamarneh, "Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?" in *Proc. of the 36th AAAI Conf. on Artificial Intelligence (AAAI)*, 2022, pp. 13 457–13 465.

[15] A. Badré and C. Pan, "Explainable multi-task learning improves the parallel estimation of polygenic risk scores for many diseases through shared genetic basis," *PLoS Comput. Biol.*, vol. 19, no. 7, p. e1011211, 2023.

[16] L. Shi, "Enhancing medical explainability in deep learning for age-related macular degeneration diagnosis," *Scientific Reports*, vol. 15, no. 16975, 2025.

[17] A. Singh, M. Agarwal, and M. Bansal, "Pose tutor: An explainable system for pose correction in the wild," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 3521–3530.

[18] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 11 802–11 812.

[19] B. Dittakavi, B. Callepalli, A. Vardhan, S. V. Desai, and V. N. Balasubramanian, "Care: Counterfactual-based algorithmic recourse for explainable pose correction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 4902–4911.

[20] G. Dibenedetto, M. Polignano, P. Lops, and G. Semeraro, "Human pose estimation for explainable corrective feedbacks in office spaces," in *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct)*. ACM, 2024.

[21] P. Contributors, "torchvision.models.detection.keypointrcnn_resnet50_fpn," https://pytorch.org/vision/stable/models/generated/torchvision.models.detection.keypointrcnn_resnet50_fpn.html, 2021, accessed: 2024-06-20.