From ‘nice try’to ‘nice throw’: exploring counterfactual explanations as corrective feedback for javelin throwing

Lennart Eing, Annika Stippler, Cristina Conati, Stefan Künzell, Elisabeth André, Silvan Mertes

Angaben zur Veröffentlichung / Publication details:

Eing, Lennart, Annika Stippler, Cristina Conati, Stefan Künzell, Elisabeth André, and Silvan Mertes. 2026. “From ‘nice try’ to ‘nice throw’: exploring counterfactual explanations as corrective feedback for javelin throwing.” In *AVI ’26: proceedings of the 2026 International Conference on Advanced Visual Interfaces, Venice, Italy, June 8-12, 2026*, edited by Antonella De Angeli, Albrecht Schmidt, Paloma Díaz, Alessandra Melonio, Niccolò Pretto, Rosella Gennari, Fabio Pittarello, María Menéndez-Blanco, and Luigi De Russis, 38. New York, NY: ACM. https://doi.org/10.1145/3811427.3811437.

From 'Nice Try' to 'Nice Throw': Exploring Counterfactual Explanations as Corrective Feedback for Javelin Throwing

Lennart Eing*
University of Augsburg
Augsburg, Germany
lennart.eing@uni-a.de

Annika Stippler*
University of Augsburg
Augsburg, Germany
annika.stippler@uni-a.de

Cristina Conati
University of British Columbia
Vancouver, BC, Canada
conati@cs.ubc.ca

Stefan Künzell
University of Augsburg
Augsburg, Germany
stefan.kuenzell@uni-a.de

Elisabeth André
University of Augsburg
Augsburg, Germany
andre@informatik.uni-augsburg.de

Silvan Mertes
Technical University of Applied Sciences Augsburg
Augsburg, Germany
silvan.mertes@tha.de

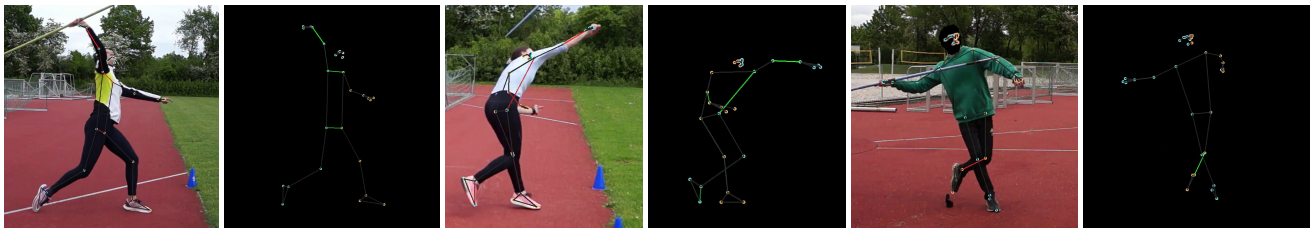

Figure 1: Athletes throwing a javelin and, on black background, counterfactuals generated using our system. Interpretations of these counterfactuals include having a more extended arm during the throw, executing a stronger follow-through, or performing a tighter cross-step.

Abstract

Providing athletes with feedback to refine their technique is key in sports coaching and is critical for improving performance and preventing injuries. However, access to expert coaching is often limited. In this paper, we explore a novel counterfactual-based feedback system as a complementary tool to expert coaching and conduct a small-scale user study to explore its perceived usability. Our approach uses an augmented GANterfactual framework, a modified CycleGAN architecture with a classifier-guided counterfactual loss, to synthesize plausible, actionable feedback. As a test bed for our approach, we use the complex motor task of javelin throwing, a sport that is characterized by high biomechanical demands and injury risk. As we are interested in the perceived usability of our approach, we conduct a user study with 21 sports students. The subjective feedback provided by participants of our user study shows that, while pose-based counterfactual feedback visualizations are appreciated by athletes, for some users they require too much domain-specific knowledge and are not "coach-like" enough. We find that athletes are looking for accompanying textual feedback, supporting recent research in the field of feedback generation for sports and motor learning.

CCS Concepts

• **Human-centered computing** → **Usability testing**; *Visual analytics*; • **Computing methodologies** → *Activity recognition and understanding*.

Keywords

Visual Feedback Systems, Explainable Sport Analytics, Motor Learning, Human-Centered Machine Learning, Counterfactual Explanations

ACM Reference Format:
Lennart Eing, Annika Stippler, Cristina Conati, Stefan Künzell, Elisabeth André, and Silvan Mertes. 2026. From 'Nice Try' to 'Nice Throw': Exploring Counterfactual Explanations as Corrective Feedback for Javelin Throwing. In *Proceedings of the 2026 International Conference on Advanced Visual Interfaces (AVI '26), June 08–12, 2026, Venice, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3811427.3811437

1 Introduction

Providing athletes with meaningful feedback to refine their technique is one of the core responsibilities of sports coaching. High-quality feedback does not only improve performance but also reduces the risk of injury, particularly in sports involving repetetive, high-intensity movements [18, 24]. However, access to expert coaching is often limited by geographic, financial, or logistical constraints. While there has been a lot of work done in recent years on methods

*Both authors contributed equally to this research.

AVI '26, Venice, Italy

ACM ISBN 979-8-4007-2342-1/26/06
https://doi.org/10.1145/3811427.3811437

for sports analysis based on video [9, 13], wearable sensors [22], and other methods, there is a lack of actionable feedback and explainability of that feedback. This can limit their usefulness in guiding athletes and coaches [3].

In this paper, we contribute to the existing literature on sports feedback systems by leveraging the GANterfactual [16] approach to generate actionable visual feedback in the form of pose-based counterfactuals, and explore the perceived usability of our approach in a small scale, exploratory user study. Instead of merely identifying important regions in a video as feature attribution methods like LIME [21] and others would, counterfactual generation approaches illustrate how small changes in the input would lead to a different outcome. This offers a direct and interpretable pathway for athletes to understand how to refine their technique.

To this end, our approach extracts structured pose data from video and transforms it into video-based counterfactual feedback. These counterfactuals are then visualized as synchronized pose videos. This allows athletes to directly compare their recorded movement to generated feedback. To ensure accessibility, the system is deployed as a web-based application that enables users to upload videos and receive feedback without the need for domain-specific expertise.

As our test-bed, we chose the track and field event of the javelin throw. It is a technically demanding sport with a high incidence of shoulder and elbow injuries due to its dynamic and repetitive throwing motion [15]. Especially during training, where the same complex throwing motion is executed over and over again, athletes can benefit from having their form corrected. To train our models, we collect a dataset of 547 videos of single javelin throw attempts at the Institute of Sports Science at the University of Augsburg.

As we are interested in the perceived usability of our approach, in addition to presenting the modifications we had to make to generate counterfactuals based on skeletal pose data using the GANterfactual approach, we conduct a small-scale exploratory user study with 21 sports students to assess the practical implications of real world use of our system and gather feedback from real users. This user study provided valuable insights into how athletes interact with generated feedback based on skeletal pose data, and helped surface concrete areas for future refinement.

To summarize, the contributions of this paper are:

- We present a novel application of classifier-guided counterfactual explanations for athletic movement based on the GANterfactual approach [16], extending prior work on static or feature-level counterfactuals to structured, temporal pose data.
- We curate a dataset of 547 annotated javelin throw videos used to train and evaluate our models.
- We conduct a small-scale exploratory user study with 21 sports student, offering real-world insight into how athletes engage with our approach. While preliminary in scope, this evaluation provides an important early step toward grounding XAI systems in real user contexts – a dimension often missing in current explainability research.

In the following, we will first give a short overview of other approaches used for similar sports analysis and feedback generation tasks in Section 2. Additionally, we give a very short introduction into the event of the javelin throw in Section 3. We then describe our data collection in Section 4, and our system in Section 5. We then describe our user study in Section 6 and discuss our results in Section 7.

2 Related Work

2.1 3D (Body) Pose Understanding

Understanding 3D body poses is often a crucial task in human activity understanding as it can be used as a low dimensional intermediary representation of stance and motion. Datasets like Humans3.6M [7] and NTU RGB-D [10, 23] have recently been used to greatly improve the understanding of human pose through human pose estimation, action classification and markerless motion capture. [1]

There are a great number of pose understanding methods recently published. Ludwig et al. [12, 13] developed a CNN-based framework for ski jumping that estimates critical flight parameters by detecting athlete ski poses as well as arbitrary keypoints on the human body. Wenninger et al. [28] evaluated the effectiveness of different machine learning approaches in classifying tactical behaviors from a dataset of 1,356 beach volleyball matches. [9] developed a new approach to infer 3D ball spin and trajectory from 2D video. This knowledge is required to be able to accurately analyze a players technique during training.

While all the methods described above provide valueable information into an athletes/team performance, they all lack actionable insights. Thus, we turn to counterfactual explanations to provide these actionable insights. We want to understand whether pose-based counterfactual explanations provide useable feedback to athletes.

2.2 Counterfactual Explanations

The use of counterfactual explanations in XAI has gained significant attention in recent years, as they pose an alternative to traditional feature importance or saliency-based methods for interpreting machine learning models. Rather than asking why a certain outcome occurred, counterfactual explanations focus on understanding what would have been needed to change for a different outcome to happen [17]. By highlighting differences between the original input and a generated counterfactual one can generate feedback in the same domain as the input data, telling a user what he would need to have done differently to improve. Existing literature provides evidence that such explanations may provide users with a more intuitive understanding of the model's decision-making process [27].

A wide range of research has explored the use of generative models to automatically generate counterfactual explanations for vision tasks. In particular, Generative Adversarial Networks (GANs) have shown great potential [8, 20, 26]. Nemirovsky et al. [20] proposed CounterGAN, a framework that produces residuals that, when added to the input image, result in a counterfactual. Their work however is tailored toward providing insight into how to improve a given algorithm, not provide feedback to the user. Khorram and Fuxin [8] proposed another framework that can generate counterfactuals by learning latent space transformations of a generative

model. However, they also do not provide any insights into user satisfaction or perceived usability with the resulting counterfactuals when used as explanations.

Another approach to generate counterfactual explanations – originally developed for the image domain – is the *GANterfactual* framework [16]. It works by integrating a classifier's decision into the loss function of a GAN to generate realistic input modifications that lead to different classification outcomes. However, all of those approaches focus on *static* images, i.e., they are not able to cope with video data. Originally applied to medical imaging, GANterfactual proved to have advantages regarding explanation satisfaction over conventional feature attribution methods like LIME and LRP, particularly with non-expert users in a variety of use-cases. For instance, [5] applied a slightly modified version of the framework to behavioral feature data in the context of job interview coaching. Their system generated feature-based counterfactual explanation from video instead of static images. These improved the perceived interviewee engagement by offering textual feedback derived from high-level behavioral metrics. In contrast, our work emphasizes direct visual feedback in the form of video-based counterfactuals to guide movement adjustments.

3 Javelin Throwing

Although our approach is applicable to a large range of different sports, we focused on the javelin throw event as an exemplary use-case. Javelin throw is an ideal testbed for counterfactual feedback, as the required biomechanical optimizations allow variations in movement to be meaningfully analyzed. Javelin throw is a track and field event where athletes aim to throw a long, spear-like implement as far as possible. According to Olympic regulations, the javelin must be thrown overhand and land tip-first within a marked sector. The athlete must remain within a designated runway and avoid crossing the foul line; otherwise, the attempt is invalid [19, 24].

The javelin throw consists of four key phases: approach run, withdrawal and cross steps, final steps and throw, and follow-through [24, 25]. In the approach run, the athlete builds speed and control while carrying the javelin above the shoulder. This phase is crucial for generating momentum that can be transferred into the throw. During the withdrawal and cross steps, the javelin is drawn back while the athlete maintains alignment with the throwing direction. These movements are essential for preserving balance and preparing the body for release. During the final steps and throw, the braced front leg plants firmly, producing a braking impulse on the body that enables the effective energy transfer from the lower extremities onto the javelin. A series of rotations lead to a delayed arm strike and maximum-force release of the javelin. The throwing arm extends fully and snaps forward to release the javelin. The follow-through allows the athlete to safely decelerate and remain balanced after the throw. It reduces the risk of injury by dissipating excess kinetic energy [25]. The different phases of the throw described above put high stresses on joints and ligaments, increasing risk of injury if improperly performed. Developing and maintaining proper technique is essential in preventing injury in javelin throw and other throwing events [2, 15].

Figure 2: Two frames taken from a raw video in the dataset, depicting the withdrawal and throw phase of a javelin throw.

4 Data Collection

We recorded a dataset at the Institute of Sports Science at the University of Augsburg consisting of a total of 547 videos of single javelin throw attempts. These attempts were performed by 70 sports students with prior training in the javelin throw. We additionally collected additional metadata on the recorded subjects like age, height, weight, handedness and gender. Attempts were performed by a total of 37 male and 33 female participants. The number of throws per person varied. Each video varies in length ranging from 5 to 30 seconds. For every attempt we recorded the achieved throw distance. Distances ranged from 8 meters to 41 meters with a mean distance of 22.46 meters and standard deviation of 6.60 meters.

The camera was positioned laterally to the runway, capturing the thrower as they ran from left to right through the cameras field of view. The recorded dataset is diverse in terms of background scenery and distance and angle of the camera relative to the thrower. Figure 2 shows exemplary video frames of the approach, withdrawal and throwing phase of an attempt. All videos were downsampled to 50 frames per second.

5 System Description and Model Training

In the following section we explain our data pre-processing procedure, counterfactual synthesis, and how feedback is presented to users in our study. It has to be noted that while we use javelin throwing as a practical example in some parts of our explanation, our approach – apart from our method to recognize the relevant video sections – can be extended to any task where: (i) There is a quantifiable measure of performance, which can be used for binary classification into "good" and "poor" performances. (ii) Feedback can be entirely based on temporal sequences of 3D pose data.

A general overview of our approach is as follows: First, 3D poses are extracted from all frames of a given input video using MediaPipe [14]. The pose data is preprocessed and normalized, ensuring uniform feature contributions. The normalized input features are fed into a classifier trained to distinguish good from poor performance. A modified CycleGAN-based generative model then synthesizes counterfactual pose sequences that transform poor performance pose sequences into improved ones, guided by the classifier. Finally, the synthesized counterfactuals are visualized and presented to users through a web application. This system enables users to compare original and improved movements in a clear and interpretable format. In the following, we describe the single steps in more detail.

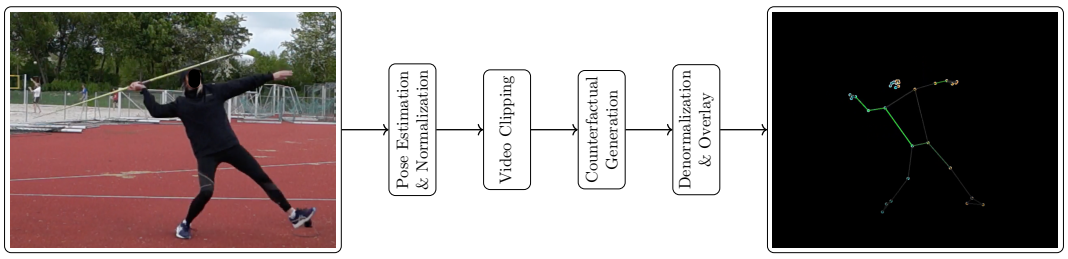

Figure 3: System Overview: For each video frame poses are estimated and normalized. Then, data is clipped to the relevant movement phases. Counterfactuals are generated, and the output is denormalized to the original input dimensions and overlayed with the pose sequences and shown to the user.

5.1 Pose Estimation and Data Pre-processing

Given the limited size of our dataset described in Section 4, maximizing the retention of relevant information is essential. However, raw video data presents high complexity, requiring any generative model working directly in the video domain to be presented with large amounts of training data. This is why it is crucial to extract structured and comparable data from the available video material. To this end, we employed the Mediapipe[1] [14] framework for full body 3D pose landmark detection, reducing the complexity of our data. The pose estimation model we used provides a total of 33 key landmarks per video frame. These include joints such as shoulders, elbows, and knees, as well as a number of facial landmarks. See Figure 4 for a visualization of the detected landmarks. Each video was processed frame-by-frame, and the extracted pose data was stored in tabular format containing the video and frame identifiers.

All 33 landmarks are normalized by image dimensions. As camera placement, angle, and athlete placement are not consistent between all videos, additional normalization was required to ensure data consistency across and within different videos. To achieve this, all coordinates were re-centered relative to the midway point between the left and right shoulders. This shifts the coordinate origin to the center of the torso, enabling the analysis of movements based on relative joint positions. After re-centering, the coordinates are scaled to $[-1, 1]$, which helps stabilize neural network training by ensuring uniform feature contributions.

To ensure that the system only focuses on information relevant to the execution of the given task, pose landmarks that are irrelevant to it were removed. In our test-bed of javelin throwing, we removed the facial landmarks (i.e., eyes, nose, and mouth) as they do not contribute to the analysis of throwing technique. We kept landmarks that describe head location and rotation. All in all, we kept a total of 22 landmarks per frame.

To clean the input data, an automated algorithm was implemented to detect the athlete's entry and the release of the javelin to isolate the relevant motion phases (approach, withdrawal, and throw). Initially, all frames without any detections at the beginning or end of the video were removed. The start of the run was identified by detecting sustained forward movement, measured by decreases in the x-coordinate of the right hip across at least 20 frames. To identify the frame corresponding to javelin release, the vertical position of the throwing hand was used the indicator, as the highest point (i.e., the maximum y value) typically marks the end of the throwing motion. It has to be noted that, while both heuristics do have an impact on the length of the pose sequence, since they **only** impact the sequence length, they do not limit the applicability of the approach to other tasks. To achieve more reliable throw detections, a small window of averaged y-coordinates is used, as it was observed that persons in the background sometimes were detected as the person of interest.

Figure 4: An example frame displaying the successfully detected pose markers inpainted onto a video.

5.2 Classifier

To generate counterfactuals, a binary classification of good and bad performance is required. To this end, we train a classifier that distinguishes good from poor performance, in our case javelin throwing attempts, using throw distance as a proxy good and bad technique. The median throwing distance (21.0 meters) is chosen to divide the dataset into good and poor performances as the threshold. We chose the median throwing distance as the cutoff threshold to ensure a balanced dataset, avoiding class imbalances during classifier training. It provides a easy to understand stand-in for more complex performance measures one could use in our approach.

[1] ai.google.dev/edge/mediapipe/solutions/

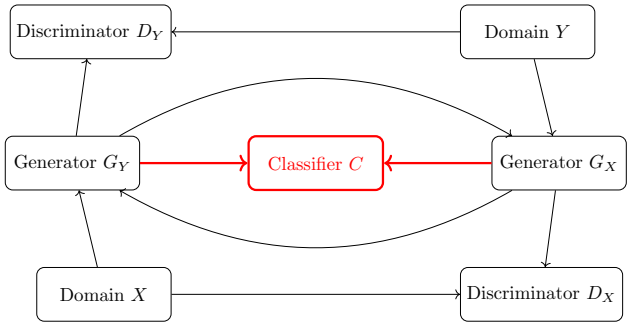

Figure 5: The GANterfactual framework [16] extends the CycleGAN framework [29] by introducing an additional optimization goal via a classifier C (highlighted in red). An additional loss component $\mathcal{L}_{counter}$ enforces generated data to flip the prediction of the classifier to the opposite class. Figure adapted from [16].

Given the small scale of our dataset, it is not a viable option to implement our classifier as a transformer model. Thus, the classifier is implemented as a LSTM [6] model. Unlike fully connected networks, the input data retains its temporal structure and is provided as a sequence of 22 3D pose landmarks flattened into a single vector with a total of 100 time steps for a total input size of 100×66.

The sequence is first processed by an LSTM layer with 50 units, outputting a full sequence of hidden states of size 100×50. This allows the model to capture temporal dependencies throughout the entire input. Then, dropout is used to drop 20% of the input units, reducing the risk of overfitting during training. The output sequence is passed through a second LSTM layer with 20 units. We use the last hidden state of the second layer as a fixed-size summary vector. This summary vector is fed into a single neuron and a sigmoid activation function, yielding a score between 0 and 1 indicating whether the input sequence corresponds to a good throw.

For training the classifier we perform 5-fold cross validation with 30 epochs using the Adam [11] optimizer. This results in a set of classifiers with an average validation accuracy of **77.88%** with a standard deviation of ±6.05%.

5.3 Counterfactual Skeleton Synthesis

We use a GAN-based architecture to transform sequences of pose data from *bad* to *good* task performances. To guide the transformation process, we use the classifier described in the previous section. Specifically, we modify a specific GANterfactual [16] generation architecture described by [4].

GANterfactual is a GAN-based framework designed to create counterfactual explanations, i.e., modified versions of the input data that change the output of a classifier while remaining realistic and interpretable. It extends the CycleGAN [29] architecture with a so-called *counterfactual loss*.

The original CycleGAN objective combines adversarial and cycle-consistency losses to enable translation between unpaired domains (e.g., bad and good throws):

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = \quad & \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ + \quad & \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ + \quad & \lambda \mathcal{L}_{\text{cycle}}(G, F) \\ + \quad & \mu \mathcal{L}_{\text{identity}}(G, F)\end{aligned} \tag{1}$$

Here, G and F are generators translating between domains X and Y; D_X, D_Y are the corresponding discriminators. $\mathcal{L}_{\text{GAN}}$ is the standard GAN loss used to encourage the generation of viable data from a given target domain. The cycle-consistency loss $\mathcal{L}_{\text{cycle}}$ ensures that translating a sample to the target domain and back reconstructs the original, encouraging transformations that are structure-preserving. Structure-preservering transformations are important, as we do not want to generate pose sequences, that are examples of good javelin throwing technique, but are too far from the original input sequence to be meaningfully actionable. Additionally, we follow a recommendation by [29] and include an *identity loss* $\mathcal{L}_{\text{identity}}$ which penalizes changing inputs that are already within the target domain.

To create meaningful counterfactuals, the GANterfactual framework extends this objective with a classifier-guided *counterfactual loss*, ensuring that generated samples not only resemble the target domain visually but are also semantically valid with respect to a trained classifier (see [16]). This is done by integrating the outputs of the frozen classifier C directly into the training process:

$$\begin{aligned}\mathcal{L}_{\text{counter}}(G, F, C) = \mathbb{E}_{x \sim p(x)} \left[\left\| C(G(x)) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\|_2^2 \right] \\ + \mathbb{E}_{y \sim p(y)} \left[\left\| C(F(y)) - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|_2^2 \right]\end{aligned} \tag{2}$$

$\mathcal{L}_{\text{counter}}$ penalizes generators G and F if the output does not shift the classifier's prediction toward the opposite class – for example, if a sample from the "bad" domain remains classified as "bad" after transformation. This encourages the generator to make semantically relevant changes to the input that are recognized by the classifier as the opposite of the input class. By doing so, it is enforced that the generated samples not only resemble the target domain, but are also classified as such. In our case, the classifier C is the distance-based javelin throw classifier introduced in Section 5.2. By this mean, we can generate pose sequences that are a minimal-change example of the input sequence, transforming bad to good throws.

The final objective then is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{cycle}} + \mu \mathcal{L}_{\text{identity}} + \gamma \mathcal{L}_{\text{counter}} \tag{3}$$

where λ, μ, and γ control the weight of cycle consistency, identity preservation, and classifier guidance losses, respectively.

As in [5], we implemented our generators as fully connected networks rather than using convolutional or recurrent layers. We adapted the generator architecture to process the 6, 600-dimensional feature vectors ($100 \times 66 = 6,600$), using fully connected layers with ReLU and tanh activation functions. As such, similar to the classifier, the counterfactual generation network processes all frames at once.

For our specific testbed of javelin throwing, we trained a set of networks on the recorded dataset (see section 4). We perform a grid search on the relevant hyperparameters of the training. We

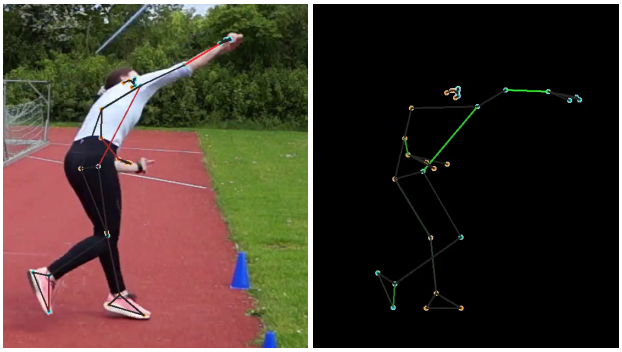

Figure 6: A frame of an analyzed sample with the original landmarks inpainted on the original (left) and the counterfactual pose (right). The counterfactual recommends the user to lean forward and extend their arm more to have stronger follow-through. We adjust the color brightness of the joint connections linearly to the deviation of counterfactual and original. As such, the user gets pinpointed to important parts of the skeleton.

keep the best two models, both of which were trained using an Adam [11] optimizer with a learning rate of 2×10^{-4} and weight decay of $\beta_1 = 0.5$. We used the weighting parameters of $\lambda = 15$, $\mu = 0.25$, and $\gamma = 0.5$.

5.4 Visual Counterfactual Feedback

To make feedback accessible without requiring technical knowledge or additional hardware, the system is deployed as a web application. Users provide a video of them performing a javelin throw via a simple web-based interface. Upon uploading poses are estimated and preprocessed, irrelevant video sections are clipped, counterfactual feedback is generated and visualizations shown to the user.

To visualize input and counterfactual pose sequences we use standard MediaPipe functionalities. The original and counterfactual sequences are clipped to the same duration and synchronized frame-by-frame. The counterfactual is displayed on a black background for clear comparison, and both videos are slowed to 0.25× speed to highlight subtle movements.

Differences between the original and counterfactual poses are quantified using 3D vector differences between landmark connections. These are mapped to color intensities: Red for the original input and green for the counterfactual feedback, with brightness of the connecting edge indicating the degree of change between them (Figure 6). This highlights which body segments were modified and to what extent.

6 User Study

To explore the usefulness of our approach, we conducted a user study to ascertain the quality of generated counterfactuals, and the comprehensibility of the corresponding feedback for real athletes.

The study was conducted at the Institute for Sport Science at the University of Augsburg. A total of 21 pre-service physical education teachers (11 female and 10 male, volunteer sampling) aged 20 years to 25 years old ($\mu = 22.9$ years, $\sigma = 1.42$ years) were recruited. They had all completed a track and field course which lasted one semester which introduced them to the javelin throw. Consequently, all participants had some proficiency in the fundamental principles of javelin throw technique. None of the participants engaged in regular training or competed at a professional level. We deliberately chose to not include a control group as we were mostly interested in the perceived usability of our approach. Utilising a control group study design would facilitate the investigation into whether our system outperforms other forms of feedback generation and presentation in terms of performance improvement. However, this was not the primary objective of this work.

Our study setup was as follows:

(1) **Warm-up:** The participant performed a thorough warm-up.
(2) **Introduction to the System:** The participant was introduced to the system using an exemplary input video. They were given a brief explanation regarding the color coding of the counterfactual and a potential interpretation of the counterfactual. All participants received the same explanation.
(3) **Demographics:** Age, gender, and weight, as well as a self-assessment of javelin throw performance of the participant was recorded.
(4) **Training Simulation** (repeated 5×):
  4a) Participants performed a javelin throw attempt. The attempt was captured using a camera.
  4b) The attempt was measured.
  4c) Counterfactual feedback was synthesized using our system using the recorded throwing attempt.
  4d) Participants watched the feedback, rewatching it as often as they wished before proceeding. This simulated how participants would use the system during actual training.
(5) **Quantitative Evaluation**: Participants were asked to rate the following questions on a scale of 1 (agree) to 5 (disagree).
  Q1: *"The system helped me become better at the javelin throw."*
  Q2: *"By using the system, I was able to further develop my javelin-throwing technique in a targeted way."*
  Q3: *"The system helped me better understand my strengths and weaknesses in the javelin throw."*
  Q4: *"Thanks to the system, I made faster progress in the javelin throw than I had expected."*
  Q5: *"I would continue to use the system to further improve my performance in the javelin throw."*
(6) **Semi-structured Interview:** Participants were asked to answer the following set of questions:
  Q1: *"What did you like about the system?"*
  Q2: *"Is the feedback in this form understandable and applicable for you?"*
  Q3: *"What did you not like about the system?"*
  Q4: *"Do you have any other comments or suggestions for improvement?"*

Answers were recorded and transcribed. In the following section participants are referred to as 'P" followed by an anonymised, numerical identifier. To be able to cite participants, transcriptions were translated from German to English.

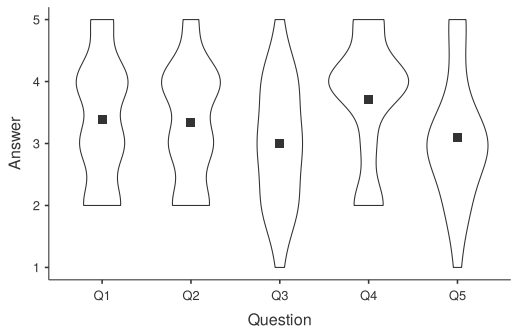

Figure 7: Distribution of the answers given by all participants given to quantitative questions Q1-Q5. The mean answer is marked as a black square.

6.1 Quantitative Evaluation

Figure 7 shows the distribution of all answers given to quantitative questions **Q1** - **Q5**. Mean answers are marked as a black square. The scale was from 1 (agree) to 5 (disagree). All responses had a standard deviation between 0.96 and 1.04. All responses had a mean value of between 3.00 and 3.71 (more specifically 3.38 for **Q1**, 3.33 for **Q2**, 3.00 for **Q3**, 3.71 for **Q4**, 3.10 for **Q5**). Overall, these results show a trend toward slight disagreement with the given questions, i.e. a slight negative perception of our system with respect to helping improve the users javelin throw technique. However, this can be explained through the answers given during our qualitative evaluation.

6.2 Qualitative Evaluation

Q1 *"What did you like about the system?"*: 20 out of 21 participants appreciated the visualization of features of the system. They found the ability to view their own throwing sequence alongside an ideal version particularly useful, as it allowed for direct comparison and immediate, actionable feedback, e.g.:

- **P3**: *"I liked that you could see your own version and the optimal version."*
- **P6**: *"That I could see myself throwing and that the mistakes were already visually represented so that I could see where they were."*

These results underscore the necessity for systems like ours that support self-assessment through visual counterfactuals that juxtapose both actual and hypothetical movements. Color-coded visualizations and the capacity to pause or replay the video in slow motion further enhanced the users' abilities to analyze and reflect on their technique, e.g.:

- **P8**: *"I liked the fact that it was essentially a direct representation of my throw and then also marked in color how good or bad the different parts were."*
- **P10**: *"I could see the technique in slow motion and analyze individual movement sequences frame by frame."*

Q2 *"Is the feedback in this form understandable and applicable for you?"*: 12 out of 21 participants described the counterfactual feedback as understandable and applicable, e.g.:

- **P21**: *"Yes, definitely, because you get immediate feedback through the video and can immediately see an improved version of how the athlete's position should have been. You can then implement this directly in your next attempt."*
- **P11**: *"I can see exactly where my weaknesses lie and I can see exactly what I need to change so that there is no longer a red line."*

Three participants found the feedback to be partially understandable and applicable, e.g.:

- **P4**: *"I would say partly yes, partly no. Because it is clearly marked, you can focus more specifically on what you need to work on. But there are also difficulties with the system. The axes overlap, and it is not always clear what is possible and what is necessary. It is particularly difficult to see this clearly when it comes to the feet."*

Out of the group of participants that found the feedback partially and not directly understandable, four participants reported difficulties in interpreting the generated feedback, e.g.:

- **P6**: *"It's good that you can see it visually, but you would then have to somehow show what exactly the error is."*

Also, four our of the 21 participants highlighted that the presented feedback may require domain knowledge to correctly derive what to change, e.g.:

- **P9**: *"Because I have previous experience with the technique and know what it should look like, I was able to interpret the suggestions, but as an uninformed test subject, I would not have known that."*

The remaining two participants did not comment on the comprehensibility of the feedback.

Q3 *"What did you not like about the system?"*: The main criticisms participants reported when interacting with the system were twofold: The identification of the actual mistake had to be performed by the user (4 out of 21 participants). Furthermore, there was no additional textual representation of what had been done incorrectly (4 out of 21 participants), e.g.:

- **P8**: *"It [the system], does not point out the mistakes directly, but only the parts of the body that are not right."*
- **P6**: *"That the error is not described in, say, text form, rather than just being marked."*

Additional recurring concerns were related to tracking accuracy (3 mentions), the limitation of available axes, i.e. only one camera angle (2 mentions), and the size of visualizations (2 mentions).

Q4 *"Do you have any other comments or suggestions for improvement?"*: participants repeatedly expressed a desire for additional textual feedback to improve the system, including a description of the mistakes they made and ideally, exercises or drills to improve their movements (8 mentions), e.g.:

- **P2**: *"So, with a keyword, catchphrase, or sentence, what should be improved?"*
- **P17**: *"I would say that figurative language should be added, such that one can understand exactly what the system wants*

you to improve. So, is it the bow tension, is it the stretched arm, ..."

7 Discussion

In our quantitative responses, we discovered that our system was slightly negatively perceived. However, we think that this can be explained through the answers given to our qualitative questions. While users generally appreciated the feedback offered by counterfactual visualizations, we found that presenting counterfactuals based on skeleton key points alone may not be sufficient for all users. Several participants reported difficulties in interpreting what exactly needed to be changed, even when there were clear visual differences. These findings highlight a limitation of purely visual counterfactuals based on skeleton key points: While they offer a direct representation of what a better movement might look like, they still require an additional interpretation step by the user.

This limitation was reflected in the recurring request for additional textual feedback. Counterfactuals based on skeleton key points seem to work well as the basis for feedback generation, but without supporting explanations, they may place too much interpretive responsibility on the user, particularly those with less experience. We think that adding even short text elements or, as one participant phrased it, *"catchphrases"* may help bridge this gap and lower the barrier to actionability. This finding is supported by work that was very recently published by Ashutosh et al. [1].

A key design challenge is related to how the system should handle already competent or high-quality throws. When the input movement is already near-optimal, the counterfactual generator – trained to transform bad throws to good throws – will still attempt to alter the input. In these cases, feedback might become unnecessary, overly subtle, or, in the worst case, biomechanically implausible, causing confusion to a user. A possible mitigation would be to include a certainty or quality score that either suppresses counterfactuals for high-quality inputs or flags them as optional suggestions rather than corrections. Integrating such awareness into the generation pipeline is an important direction for future work.

A more fundamental challenge lies in how classifier errors affect the training of the counterfactual generation model itself. Since our generation model is guided by a classifier that defines what constitutes a *good* versus *bad* throw, misclassifications during training can introduce noise into the training process of the generator. For instance, if a poor-quality throw is mistakenly labeled as *good* (a false positive), the generator may learn to emulate flawed techniques, embedding suboptimal patterns into the target domain. Conversely, false negatives - *good* throws mislabeled as *bad* - can distort the generator's understanding of what to correct, leading to unnecessary or counterproductive modifications. Such issues not only degrade feedback quality but may also undermine the model's ability to converge on physiologically sound transformations. More reliable classification thresholds or even joint training strategies that account for label uncertainty may help mitigate these risks.

One last limitation of our work lies in the selection of the user group selected for our exploratory study. Our results are currently limited to a group of users that is already familiar with the basics of the javelin throw, while none of them are throwing on a professional level.

8 Conclusion & Outlook

In this paper, we presented a system for generating counterfactual visual feedback to support athletes in refining their technique using the GANterfactual framework. We demonstrated our approach in the context of javelin throwing. By combining pose estimation, a classifier-guided GAN, and a web-based interface to present them, our method provides intuitive, movement-based feedback.

The results of our study provide initial support for the feasibility and perceived usefulness of the proposed counterfactual feedback system. Participants largely responded positively to the visual feedback features, particularly the ability to compare their own movements with a corrected version of their movement. The idea of showing a plausible alternative improved version of the movement was understood by most participants and helped them identify areas in need of adjustment. We also identified the main area for improvement in an additional layer of actionable feedback, namely textual feedback.

For future work, an emphasis should be put on the generation of multimodal counterfactual feedback, classifier safeguarding, and a wider range of study participants (beginners, athletes, coaches, etc.). In our study, participants repeatedly mentioned the wish for a textual recommendations of what should have been changed to increase the comprehensibility of the feedback. Generation of corresponding textual counterfactual feedback could be realized by employing an LLM.

Additionally, future work should investigate how AI-driven systems can effectively enhance human coaching practices. Understanding how coaches and athletes interpret and act on counterfactual feedback could help integrate these tools more effectively into real training environments. Although only tested on javelin throwing, our approach might still be broadly applicable to other sports that involve complex, technique-dependent actions such as long jump, baseball pitching, or gymnastics. Applying our method across a wider variety of sports will help evaluate its generalizability and study domain-specific considerations for feedback design.

Acknowledgments

This work was funded in parts by the BIGEKO project (BMBF, German Ministry for Education and Research, grant number 16SV9094), TherapAI project (DFG, German Research Foundation, grant number 493169211), and FORSocialRobots project (BFS, Bavarian Research Foundation, grant number AZ1594-23).

References

[1] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. 2025. ExpertAF: Expert Actionable Feedback from Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13582–13594.

[2] K. Bartonietz and A. Bartonietz. 1995. The throwing events at the World Championships in Athletics 1995, Goteborg - Technique of the world's best athletes, Part 3: Javelin throw. *New Studies in Athletics* 10 (1995).

[3] Vanessa Camilleri, Reno Yuri Camilleri, Mark Fialovszky, Daniel Pace, Dylan Seychell, and Matthew Montebello. 2025. Towards Explainable Multimodal Sensing for Swimming Analysis: Early Findings from the SWIM-360 Project. *Sensors (Basel, Switzerland)* 25, 22 (November 2025), 7047. doi:10.3390/s25227047

[4] Alexander Heimerl, Silvan Mertes, Tanja Schneeberger, Tobias Baur, Ailin Liu, Linda Becker, Nicolas Rohleder, Patrick Gebhard, and Elisabeth André. 2022.

Generating personalized behavioral feedback for a virtual job interview training system through adversarial learning. In *International Conference on Artificial Intelligence in Education*. Springer, 679–684.

[5] Alexander Heimerl, Silvan Mertes, Tanja Schneeberger, Tobias Baur, Ailin Liu, Linda Becker, Nicolas Rohleder, Patrick Gebhard, and Elisabeth André. 2022. "GAN I hire you?" – A System for Personalized Virtual Job Interview Training. arXiv:2206.03869 [cs.HC] https://arxiv.org/abs/2206.03869

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

[8] Saeed Khorram and Li Fuxin. 2022. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10203–10212.

[9] Daniel Kienzle, Robin Schön, Rainer Lienhart, and Shin'Ichi Satoh. 2025. Towards Ball Spin and Trajectory Analysis in Table Tennis Broadcast Videos via Physically Grounded Synthetic-to-Real Transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5842–5851.

[10] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2020. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2684–2701.

[11] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[12] Katja Ludwig, Moritz Einfalt, and Rainer Lienhart. 2021. Robust Estimation of Flight Parameters for Ski Jumpers. In *Proceedings of the Multimedia Computing and Computer Vision Lab, University of Augsburg*. University of Augsburg.

[13] Katja Ludwig, Daniel Kienzle, Julian Lorenz, and Rainer Lienhart. 2023. Detecting arbitrary keypoints on limbs and skis with sparse partly correct segmentation masks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 461–470.

[14] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC] https://arxiv.org/abs/1906.08172

[15] Adele Meron and Deborah Saint-Phard. 2017. Track and Field Throwing Sports: Injuries and Prevention. *Current Sports Medicine Reports* 16, 6 (November/December 2017), 391–396. https://journals.lww.com/acsm-csmr/fulltext/2017/11000/Track_and_Field_Throwing_Sports__Injuries_and.8.aspx Accessed: 2024-07-22.

[16] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. 2022. GANterfactual-Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence* 5 (April 2022). doi:10.3389/frai.2022.825565 Accessed: 2024-07-20.

[17] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[18] Gregory D. Myer, Benjamin W. Stroube, Christopher A. DiCesare, Jensen L. Brent, Kevin R. Ford, Robert S. Heidt Jr, and Timothy E. Hewett. 2013. Augmented Feedback Supports Skill Transfer and Reduces High-Risk Injury Landing Mechanics. *The American Journal of Sports Medicine* (Mar 2013).

[19] Utathya Nag. 2022. Javelin throw: Know the rules, scoring system and competition format. https://olympics.com/en/news/javelin-throw-rules-regulations-and-all-you-need-to-know Accessed: 2024-07-27.

[20] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. 2020. CounteRGAN: Generating Realistic Counterfactuals with Residual Generative Adversarial Nets. *arXiv preprint arXiv:2009.05199* (2020).

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

[22] Minwoo Seong, Gwangbin Kim, Yumin Kang, Junhyuk Jang, Joseph DelPreto, and SeungJun Kim. 2024. *Counterfactual Explanation-Based Badminton Motion Guidance Generation Using Wearable Sensors*. doi:10.48550/arXiv.2405.11802

[23] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.

[24] Heiko K. Strüder, Ulrich Jonath, and Kai Scholz. 2023. *Track & Field Training & Movement Science - Theory and Practice for All Disciplines*. Meyer & Meyer Sport.

[25] Juris Terauds. 1985. *Biomechanics of the Javelin Throw*. Delmar, California. 53–59 pages.

[26] Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, and Oliver Cobb. 2021. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123* (2021).

[27] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018). doi:10.2139/ssrn.3063289

[28] Sebastian Wenninger, Daniel Link, and Martin Lames. 2020. Performance of machine learning models in application to beach volleyball data. *International Journal of Computer Science in Sport* 19, 1 (2020), 20–30. doi:10.2478/ijcss-2020-0002

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.