VisMimic: Integrating Motion Chain in Feedback Video Generation for Motor Coaching

Liqi Cheng
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
lycheecheng@zju.edu.cn

Xiao Xie*
Department of Sports Science
Zhejiang University
Hangzhou, Zhejiang, China
xxie@zju.edu.cn

Yiwei Peng
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
yiweipeng.cs@gmail.com

Minghao Feng
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
minghaofeng@zju.edu.cn

Yuchen He
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
heyuchen@zju.edu.cn

Anqi Cao
Department of Sports Science
Zhejiang University
Hangzhou, Zhejiang, China
caoanqi@zju.edu.cn

Yihong Wu
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
wuyihong@zju.edu.cn

Hui Zhang
Department of Sports Science
Zhejiang University
Hangzhou, Zhejiang Province, China
zhang_hui@zju.edu.cn

Yingcai Wu
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, Zhejiang, China
ycwu@zju.edu.cn

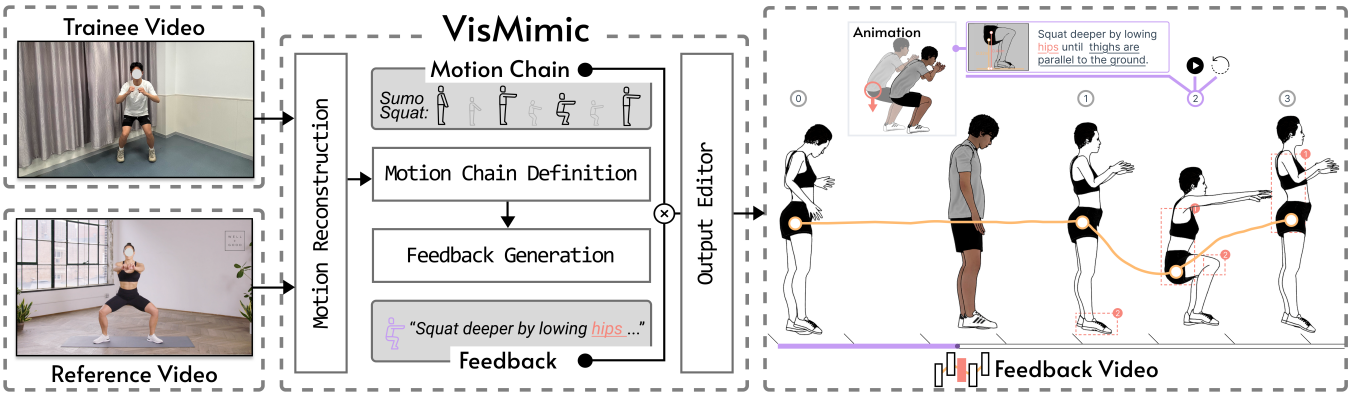

Figure 1: VisMimic takes raw video footage from both the trainee and the reference as input and outputs an augmented feedback video. To achieve this, VisMimic takes following steps: 1) reconstructing 3D human motion from the videos; 2) transforming motion data into motion chain structure; 3) generating feedback in the form of textual instructions and visual animations; 4) supporting editing of the visual representations and observation perspectives.

Abstract

Augmented video is a common medium for remote sports coaching, facilitating communication between trainees and coaches. Existing video augmentation techniques struggle to simultaneously convey both the overall motion dynamics and static key poses. This limitation hinders feedback comprehension in motor learning, making it difficult to understand where errors occur and how to correct them. To address this, we first reviewed popular video augmentation solutions. In collaboration with professional coaches, we integrated motion chain into feedback videos to combine key poses with motion trajectories. It supports multi-view observation and feedback explanation from overview to detail. To assist coaches in creating feedback videos, we present VisMimic, a human-AI interaction system that automatically analyzes trainee videos against reference movements, generates animated feedback, and enables customization. User studies show VisMimic's usability and effectiveness in enhancing motion analysis and communication for motor coaching.

*Corresponding author: Xiao Xie, xxie@zju.edu.cn

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747794

CCS Concepts

• **Human-centered computing** → **Visualization application domains**; **Visualization**.

Keywords

Augmented Sports Video, Motion Correction, Feedback Generation

ACM Reference Format:
Liqi Cheng, Xiao Xie, Yiwei Peng, Minghao Feng, Yuchen He, Anqi Cao, Yihong Wu, Hui Zhang, and Yingcai Wu. 2025. VisMimic: Integrating Motion Chain in Feedback Video Generation for Motor Coaching. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3746059.3747794

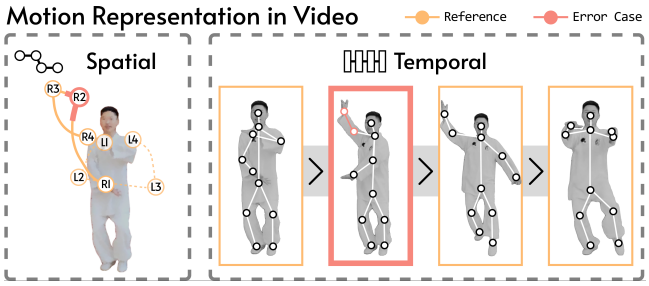

Figure 2: Current feedback presentation can be categorized along two critical dimensions: spatial and temporal.

1 Introduction

In motor skill learning, video plays a pivotal role in facilitating communication between coaches and learners, particularly when in-person instruction is not feasible. In a typical video-based coaching scenario, learners record their movements and send the footage to coaches, who drawing on their expertise, provide two types of feedback: 1) corrective instructions targeting movement errors, and 2) reference videos demonstrating the correct technique. While this dual-channel feedback enhances information flow, it also introduces an inherent gap in learners' understanding of movements:

First, there are inherent difficulties in integrating textual feedback with visual demonstrations. Textual feedback from coaches, such as "raise your elbow," requires learners to mentally translate language into movement, whereas demonstration videos rely on visual-to-motor imitation. Without a clear connection between the two, learners may struggle to associate abstract feedback with specific movement adjustments. **Second, individual differences can lead learners to interpret the same textual instruction in different ways.** Differences in physical ability and experience may cause them to perform the same instruction differently. For example, "slightly bend your knees" might lead one learner to bend 30 degrees and another to bend 60 degrees, making training results less consistent and effective. Therefore, a key to enhancing coach–learner communication is the seamless integration of coach proposed feedback into video content.

Video augmentation has emerged as a widely recognized approach for bridging textual and visual information [85]. However, current methods still face challenges across the two stages of motion feedback. In the **Feedback Generation**, text-based instructions depend heavily on the coach's expertise. When reviewing a learner's performance, coaches must stay involved throughout the entire process: observing the movement, analyzing motion data, generating corrective suggestions (text), and embedding these corrections into videos through techniques such as visual highlighting [22] and linking [31]. In the **Feedback Presentation**, effective motion understanding requires support for both: 1) Precise interpretation of key postures at specific time points, and 2) Comprehensive assessment of continuous movement dynamics. To meet these needs, video augmentation must integrate feedback across two critical dimensions: 1) **Spatial**: identifying the erroneous body part and guiding how to correct it; 2) **Temporal**: locating the key frame where the error occurs and illustrating the subsequent right movement to follow. Based on our preliminary study, we classify existing methods into two categories (Fig. 2): 1) dynamic trajectories that enhance spatial understanding [60, 75], and 2) sequential snapshots that represent temporal patterns [28, 69]. Nevertheless, current feedback presentation approaches often focus on only a single dimension of feedback, limiting the clarity of the feedback delivered. Thus, coaches need to handle both feedback creation and delivery.

To address these challenges, we introduce a semantic understanding workflow to assist coaches in feedback video creation. For **Feedback Generation**, a motion understanding model [24] extracts structured textual feedback by comparing trainee motion against a reference, enabling semantic translation from motion differences to language. Meanwhile, a text-driven motion editing model [16] transforms these textual instructions into guidance animation by adjusting joint trajectories or motion magnitudes, thereby supporting knowledge transfer from language to visual representation. This workflow alleviates the cognitive and labor burdens on coaches. For **Feedback Presentation**, we propose a representation that integrates the motion chain as its underlying data structure to convey both the overall motion dynamics and static key poses. The motion chain, a widely used concept in motion analysis [30, 78], consists of a sequence of interconnected key poses and their relationships. Inspired by the character animation editing interface [2, 84], the resulting augmented video combines key poses with motion trajectories, supporting both critical dimensions of motion feedback: 1) Spatial — enabling in-depth observation of static key poses, and 2) Temporal — providing a complete overview of continuous motion trajectories.

Designed for coaches as the target users, VisMimic was built on a consolidated workflow to generate augmented feedback videos from trainee and reference footage for motor coaching. VisMimic employs state-of-the-art video motion capture models [51] to reconstruct 3D human motion and full-body avatars, enabling motion comparison. Through semi-automated key pose extraction and kinematic constraint setting, the system organizes motion data into a structured motion chain representation, which supports detailed analysis and the generation of feedback candidates. Our novelty lies in the integration of the motion chain representation and its step-by-step creation workflow. This unified approach enables coaches to incorporate experiential knowledge, perform detailed analysis, and iteratively improve feedback quality. VisMimic demonstrates practical applicability and lays the foundation for more user-friendly end-to-end AI coaching tools by: (1) defining a practical input-to-output feedback pipeline, (2) enabling controllable editing, and (3) supporting domain knowledge integration via structured data handles. In summary, our research has three main contributions:

- a structured representation , which integrated motion chain in feedback videos to combine key poses with motion trajectory, offers a foundation for human-AI collaboration;
- a system, VisMimic, that supports motion analysis for feedback generation and video creation;
- user studies that validate the efficiency of generated feedback video from both coaches and trainees perspective and utility of VisMimic for motor coaching.

2 Related Work

In this section, we review relevant research, focusing on human motion data visualization and human motion analysis.

2.1 Data Visualization for Motor Coaching

Data visualization plays a crucial role in motor coaching. It can help users understand motion data that consists of continuous human poses with diverse spatial and temporal variations [34]. To enhance coaching effectiveness, it is primarily applied in two ways: representing motion and augmenting feedback.

Motion Representation. Visualizing motion capture data is essential for motion pattern recognition and analysis. To encode the sequential nature, various visualization approaches have been proposed. For example, Motion Belts [80] and Motion Volume [52] display motion capture data as short clips of selected key frames, providing an abstract illustration of motion. OutFlow [72] employs hierarchical clustering to reduce the number of poses in flow visualization, and generates video clips associated with camera paths for effective motion overviews [12]. Another category of visualization aims to characterize relationships among different poses in a motion database [27], often displayed in a 2D space [34]. Pretorius et al. [55] proposed a bar tree to abstract multivariate motion state transitions, while Blaas et al. [17] aggregated transitions into spline bundles in a 2D graph layout. These methods provide both an overview and detailed insights into specific motion transition patterns [58]. Additionally, PoseCoach [46] and MotionFlow [36] enable the comparison of motions. However, these methods encounter certain limitations. Since motion data has semantic meaning, coaches often need to understand the context and intent behind specific movements, which can be obscured in abstract representations. Moreover, the complexity of motion data can lead to information overload. Thus, feedback augmentation methods are introduced to illustrate coaching feedback to users directly.

Feedback Augmentation. Data visualization techniques have been explored to enhance feedback in motor coaching, with approaches ranging from 2D to 3D. 2D augmented video is a common medium for sports coaching. For instance, MotionPro [5] visualizes baseball and golf swing trajectories in videos. Semeraro et al. [60] investigate visual cues to enhance movement learning in instructional videos. Clarke et al. [22] adapt the playback speed based on user motion to support real-time alignment. Tang et al. [65] combines skeleton overlays and corrective feedback for physiotherapy guidance. However, 2D augmentations face limitations such as depth ambiguity and restricted spatial context due to the 3D nature of motion. Recently, 3D approaches leveraging AR/VR have shown promise in providing immersive feedback for sports coaching. Wu et al. [75] explored AR environments for at-home workouts, offering real-time visual feedback on users' movements. Lin et al. [45] developed an AR basketball free-throw training system, providing visual feedback on shot trajectories alongside ideal paths. Similarly, avaTTAR [50] combines on-body and detached AR cues to refine table tennis strokes. To help users apply augmented feedback in training, Video2MR [35] demonstrates the potential to generate 3D instructions from videos, while Augmented Coach [71] annotates 3D volumetric recordings to critique athletic form. Although 3D augmentation solutions address spatial challenges, their reliance on head-mounted displays (HMDs) limits accessibility compared to video content that offers broader compatibility and flexibility.

Thus, we integrate 3D information of motion capture data into augmented videos, which are widely used in motor coaching [60], to provide precise feedback and support multi-view observation for users. Specifically, we aim to employ a user-centered approach to explore potential improvements by combining key poses with kinematic information (e.g., trajectory) for feedback representation.

2.2 Models for Human Motion Analysis

Human motion analysis is essential in motor coaching, providing feedback to improve performance and prevent injuries. Video is an accessible method for analyzing human motion and is widely used by coaches and trainees. Traditional video-based approaches mostly rely on manual annotation to extract motion features (e.g., joint angles, positions) for comparison with standard motions [46]. Recent advances in pose estimation enable direct extraction of 2D/3D motion capture data from videos, supporting automated pose comparison and analysis. For example, AI Coach [69] and AIFit [29] apply deep learning for personalized training analysis, while Pose Tutor [25] and 3D Pose Based Feedback [83] focus on providing pose corrections. However, most existing methods emphasize static key poses [26], overlooking dynamic kinematic characteristics in continuous motion, and often provide scores or simple comparisons without actionable or contextualized guidance for trainees [14].

Recent efforts have integrated natural language in human motion analysis. Datasets such as PoseScript [23], FLAG3D [66], MotionBank [77], and ActivityNet [33] bridge motion and language by providing natural language annotations of motion instructions, facilitating the generation of actionable feedback. For example, MotionGPT [37] treats human motion as a foreign language, translating movement patterns into semantic representations to support multiple motion tasks. Other approaches, such as PoseFix [24], Mojito [61], MotionLLM [20], and ChatMotion [43], further explore text-driven motion correction and feedback generation. However, these methods are often limited to basic analysis, largely due to the absence of domain-specific assessment criteria and fine-grained instruction-tuning data. This restricts their ability to provide holistic and task-specific feedback for motor coaching.

Recent studies have demonstrated the feasibility of bridging human motion and action semantics through kinematic phrase representations to support motor coaching by domain-specific assessment criteria [48]. Thus, our work aims to integrate automated motion analysis with coach-driven insights to facilitate feedback delivery that is both actionable and comprehensible to trainees.

3 Background

In this section, we first introduce the concept of feedback videos in the context of coaching, followed by the motion chain, a common representation used to analyze human motion patterns.

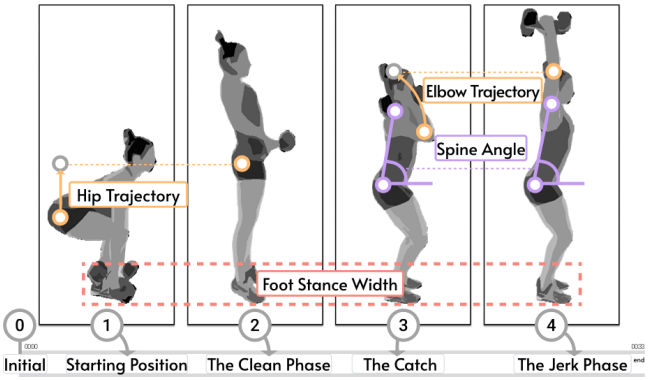

Figure 3: An example of decomposing *"the dumbbell power clean and push jerk"* into a motion chain structure.

Feedback Video. Feedback videos are post-exercise recordings enriched with personalized guidance or corrections based on a trainee's performance [13]. Their growing popularity stems from their convenience, accessibility, privacy protection, and the rise of at-home workouts. Unlike generic instructional videos, feedback videos are tailored to individuals, focusing on specific motion segments that require improvement. Coaches review the trainee's recordings, analyze errors, and insert comments or visual augmentations (e.g., visual cues) at key moments to clarify feedback [54]. Applications such as SwingVision [8], VisualEyes [9], and Volt Athletics [10] support this process by allowing annotation, motion playback, and side-by-side comparison with expert demonstrations. Effective feedback videos should clearly indicate **(1) what motion errors exist compared to the reference standard** and **(2) how to correct the error and perform the movement properly**.

Motion Chain. Motion is a sequence of continuous pose changes that describe the trajectory of human poses over time. A pose is the fundamental unit of motion, representing the joint positions and angles of the body at a specific time point. It corresponds to a static frame within a dynamic movement. Thus, motion can be represented and analyzed using a set of key poses. Based on temporal relationships, key poses are categorized into four coarse-grained types: initial pose, extreme pose, transition pose, and final posel, following prior work [64]. A motion sequence may include multiple extreme and transition poses, each carrying semantic significance. These poses are connected through kinematic constraints,including joint trajectories as well as angles and distances between multiple joints. Thus, we define a motion chain as a sequence of interconnected key poses and their relationships [30, 78].

For example, as shown in Figure 3, the motion chain of *"the dumbbell power clean and push jerk"* consists of four key poses based on movement phases and domain knowledge, excluding the initial pose (p_0). These key poses are: the starting position (p_1), the clean phase (p_2), the catch (p_3), and the jerk phase (p_4). The remaining frames consist of transition poses, representing intermediate movements between key poses. The entire motion is represented as: $M = [p_0, p_1, p_2, p_3, p_4]$. We outline several key relationships between the key poses: *Temporal Order*: The clean phase precedes the jerk phase. *Kinematic Constraints*: During the clean, the body must extend the hips to stand upright (p_1-p_2). During the jerk, the arms must lift the dumbbell overhead (p_3-p_4). Throughout the motion, it is crucial to keep the feet stable and the spine upright.

However, current feedback videos often lack the granularity needed to highlight motion patterns, leading to suboptimal coaching outcomes. Our goal is to identify space to enhance feedback videos for more effective support in video-based coaching.

4 Preliminary Study

In this section, we conducted a preliminary study to investigate the challenges faced in current video-based coaching and to uncover potential improvements for enhancing feedback support.

4.1 Procedures

Our study follows a user-centered process, including a review of existing feedback representations and interviews with professional coaches to understand their feedback generation practices.

4.1.1 Literature Review. To investigate existing video augmentation approaches for representing human motion with the feedback, we first conducted a literature review covering 20 research studies [12, 22, 28, 34, 36, 38, 39, 42, 44, 46, 50, 52, 59, 60, 65, 67, 69, 71, 75, 80], 10 popular commercial sports training applications [1, 3–11], and the 20 most viewed English-language workout videos on YouTube that include augmented visual elements. After identifying research studies and applications sources via an initial Google keyword search filtered for motion representation relevance, two authors independently applied inclusion/exclusion criteria to ensure quality. Sources were excluded if they: (1) contained no or very few motion augmentations, (2) were not primarily focused on representing human motion, or (3) did not depict a complete sports motion. The resulting sources show the diverse design space for augmented motion data visualizations. Through this screening, we observed existing representations predominantly focused on spatial or temporal aspects.

Building on this observation and prior work [79] analyzing spatio-temporal characteristics of motion, we developed a codebook through an iterative process [85] centered on **temporal alignment** as the key distinguishing factor – specifically, **whether augmentation elements evolved dynamically and continuously in real-time with the performer's movement**. An initial draft codebook was refined via preliminary analysis of a subset of sources: Three authors independently applied the draft criteria, discussed discrepancies, and iteratively refined classification rules until consensus was reached. Using the finalized codebook, the authors conducted independent manual coding. To enhance validity and minimize bias, classifications were further reviewed in consultation with two external experts, including a senior sports coach with over a decade of experience advising national teams. Disagreements were resolved through iterative discussion, ensuring consistency. Through this dual-review process, we identified two predominant types of motion augmentation representations: (1) dynamic trajectories and (2) sequential snapshots. As shown in Fig. 4.

- **Dynamic Trajectory:** This approach represents continuous motion by visualizing joint or full-body movements over time (Fig. 4.A). Common designs include directional arrows to indicate movement direction, speed, and distance, as well as body highlights and visual metaphors that update in real time. These augmentations help users understand how to correct motion by visually following the ideal path.

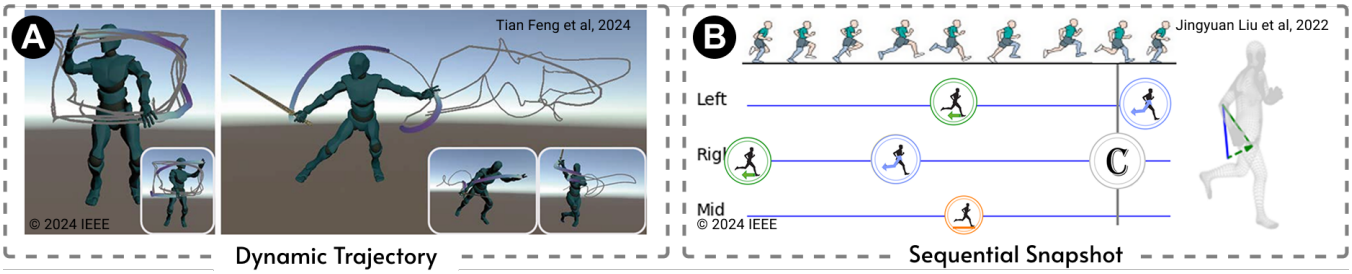

Figure 4: We classify existing video augmentation methods for representing human motion based on temporal alignment into two categories: Dynamic Trajectory [67] and Sequential Snapshot [46] (Reprinted with permission from IEEE Transactions on Visualization and Computer Graphics, Copyright of © 2024 IEEE).

- **Sequential Snapshot:** This method decomposes motion into discrete static poses, each augmented with visual elements such as guide lines and reference planes (Fig. 4.B). It is useful for analyzing motion pattern and identifying key errors at specific stages. By presenting motion in a step-by-step format, users can engage in targeted learning based on individual poses.

4.1.2 Interview. Building on the previous studies and our initial classification, we conducted semi-structured interviews with eight professional sports coaches (E1-E8; Age: 26-37), each with at least three years of experience in video-based coaching.

We first collected background information about their remote coaching practices, focusing on how feedback videos were used to deliver instructions and their feedback generation practices. Participants then viewed three sets of feedback video clips illustrating different exercises: two foundational workout movements (Hip Bridge and Push-ups) and one complex Tai Chi movement ("White Crane Spreads Its Wings"). Each movement was presented in three augmentation formats, following design guidelines from Semeraro et al. [60] and Wu et al. [75]: (1) text-only, (2) dynamic trajectory, and (3) sequential snapshots. Each clip lasts about 30 seconds.

After viewing each set, participants were asked to select the most effective format and explain why. Through open-ended discussion, they compared the strengths and limitations of each approach. From the interview transcripts, we extracted key challenges faced in using feedback videos for instruction. We also introduced the concept of the motion chain, which is commonly used in motion analysis, and presented two example cases. Coaches E1-E5 were already familiar with this concept. Participants were encouraged to suggest possible improvements to current augmentation methods. Each interview lasted approximately 25 minutes.

4.2 Findings and Discussions

4.2.1 Challenges Faced in Video-Based Coaching. While video is commonly used to support feedback in motor skill coaching, several challenges persist on feedback generation and representaion.

- **C1: Existing motion representation struggle to simultaneously convey both the overall motion dynamics and static key poses.** Most current augmentation methods emphasize either dynamic movement (e.g., motion trajectories) or static posture (e.g., key poses), but not both. As E1 explained, *"The motion trail gives a good sense of how the body moves through space, but it lacks detail at specific poses. On the other hand, snapshot visuals show individual positions clearly, but I lose the sense of how the motion flows."* Coaches emphasized the need for a feedback representation that supports both the understanding of movement flow and the recognition of critical moments within motion.
- **C2: Current feedback representation is insufficient for guiding motion correction.** Most existing feedback videos support two core tasks: identification (showing what the correct movement looks like) and comparison (highlighting deviations from the ideal form). As E3 noted, *"I can easily identify the error and recognize the gaps from the augmented video."* However, they fall short in supporting correction—helping trainees understand how to adjust their motion. This limitation places a high cognitive load on trainees, especially when dealing with unfamiliar or complex movements. E1 and E5 highlighted this challenge, particularly with complex movements: *"These feedback formats alone are not sufficient for coaching; a corrective motion demonstration may be necessary for trainees to follow and adjust accordingly"*.
- **C3: Creating feedback videos is both challenging and time-consuming.** Visual guidance has been shown to effectively convey feedback [31, 60], yet existing feedback videos often lack such enhancements. In our interviews, four coaches stated they "never or seldom use visual guidance for feedback". Despite available design guidelines [75], many coaches lack the visual design skills needed to produce high-quality feedback videos, making the process both difficult and time-intensive. Seven out of eight coaches emphasized the complexity and effort required to generate effective augmented videos. Notably, most participants had little or no prior experience with video-generation tools, indicating that the challenge stems not only from design complexity but also from limited exposure to or training in such tools.

4.2.2 Potential Improvements for Video-Based Coaching. Based on participants' suggestions for enhancing feedback videos, we identified two core requirements to address the challenges observed:

- **Feedback Representation: Combing static key poses with motion dynamics into feedback video.** The current feedback video displays movement trajectories and key poses separately, which hinders comprehension in motor learning. We propose integrating motion chain structure to address this issue. By combining the two elements, we can provide an overview of the movement pattern through key poses while also showing the movement details at each step (C1). Additionally, using key poses as nodes on a timeline allows for interactive switching between multiple views to observe and demonstrate animations from error key poses to the correct ones.

- **Feedback Generation: Supporting easy-to-use feedback video generation.** Augmented visualization is effective for understanding [85] and comparing information [75], and has been proven to be a valuable method in motor coaching [45]. The creation of augmented feedback videos relies on the coach's abilities in motion analysis and visual design. Our goal is to utilize 3D motion reconstruction technology to compare the trainee's movements with a reference standard, thereby assisting coaches in providing feedback. Additionally, based on the design space of augmented visualizations for movement [60], we aim to provide feedback through animations and augmented visualizations to make corrective feedback more comprehensible.

These requirements guided the design of VisMimic: 1) Integrating motion chains into feedback videos to combine key poses with motion trajectories (C1); 2) Utilizing state-of-the-art motion analysis models to support coaches in providing feedback (C3); 3) Augmenting feedback with correctional animations that visually demonstrate the transition from incorrect to correct movement (C2).

5 Feedback Video with Motion Chain

Inspired by character animation interfaces [2, 84], we propose an integrated motion chain representation for feedback videos. We first introduce the overview, then the underlying motion chain structure, and finally the VisMimic workflow built upon it.

5.1 Representation Overview

To combine key pose with trajectory, this representation consists of following components: trainee motion, reference key poses, transition trajectories, feedback annotations and supportive UI elements. It is screen-friendly, supporting zooming and playback controls for flexible viewing and interaction. Inspired by TimeTunnel [84], the view design multiplexes time and space. For cases where people remain in place, the reference/timeline remain static while trainee motion and trajectories advance Fig. 5. Alternatively, this representation can shift the reference dynamically while keeping the trainee spatially centered.

- **Trainee Motion** serves as the primary content in the feedback video, presenting the movements that require correction. It is positioned as a motion progress indicator, where dynamic movements and motion errors are displayed through playback control.
- **Reference Key Poses** are reference models placed alongside the trainee's motion to illustrate correct poses at key moments. They are aligned with the trainee's key pose timeline. With highlighted error body parts and feedback annotations, users can easily locate when and where motion errors occur. Each key pose serves as an interactive node. It supports multi-view rotation and playback to show correct dynamic movements and enables users to navigate quickly and access detailed feedback.
- **Transition Trajectories** are lines connecting adjacent reference key poses, illustrating the reference motion flow between them. These trajectories indicate temporal progression and the spatial relationship between key poses. As shown in Fig. 5, the trajectory between key poses are linked through related body parts with movement. Aligned with the timeline, these trajectories allow users to compare key joint differences between the reference motion and the trainee's current pose.
- **Feedback** is delivered in two formats. Feedback cards provide a snapshot highlighting the differences from the reference standard to visually indicate the motion error. Feedback animations transform textual feedback into corrective animations, directly demonstrating what to do to correct the error, helping users intuitively understand the required adjustments.
- **UI Elements** includes a timeline with navigation and playback controls. The timeline is scalable, and its orientation aligns with the perspective — side view along the X-axis and front view along the Y-axis (as shown in Fig. 12). It also supports dynamic and static reference positioning.

5.2 Motion Chain Structure

The motion chain is the underlying data structure that supports aforementioned representation.

5.2.1 Key Pose. Key poses are a set of representative poses that concisely capture the essential structure of a full-body movement [15]. The use of key poses has been widely applied in video summarization, motion retrieval [52], and character animation [84]. In animation or video indexing, key poses are often extracted by sampling frames at a fixed interval. However, in motor skill learning, each key pose carries unique semantic meaning within the specific context of the movement. Following the scope defined in prior work [64] and incorporating expert recommendations, we identify key poses based on their functional importance in the movement, extract the corresponding frames, and arrange them sequentially according to their temporal order.

5.2.2 Kinematic Constraint. This section introduces the relationship between key poses (kinematic constraints) we study. Based on prior research in sports biomechanics [46], we initially classify key pose constraints into **angular**, **positional**, and **temporal** categories. However, relying solely on this coarse-grained classification still requires significant manual effort to extract user-defined pose attributes from motion capture data. While it is straightforward for single joints, handling multi-joint relationships (e.g., distance or contact detection) is more complex. Inspired by work linking 3D human poses and natural language, we employ the concept of PoseCode [23] to enhance the constraints, which defines relationships between specific joint sets in key poses. To effectively represent kinematic constraints, we further refine angular and positional constraints by incorporating elementary relations: angle, pitch, and roll for angular constraints, and distance and relative position for positional constraints. In the following we detail these kinematic constraint classifications.

Angular. Angular constraints represent the orientation of a body segment or line relative to a reference joint. These constraints are defined by a triplet representing the first side endpoint, the vertex, and the second side endpoint. The vertex is a joint, and the endpoints can be both joints or a joint and an axis.

- **Angle** describe the bending of a body part, defined by three key joints (e.g., the angle between the upper arm and torso is formed by the elbow, shoulder, and hip).
- **Pitch & Roll** describe the verticality or horizontality of a body part, defined by two key joints and a relative axis (e.g., the neck and hip, along with the hip's z-axis, define the spine's bend angle).

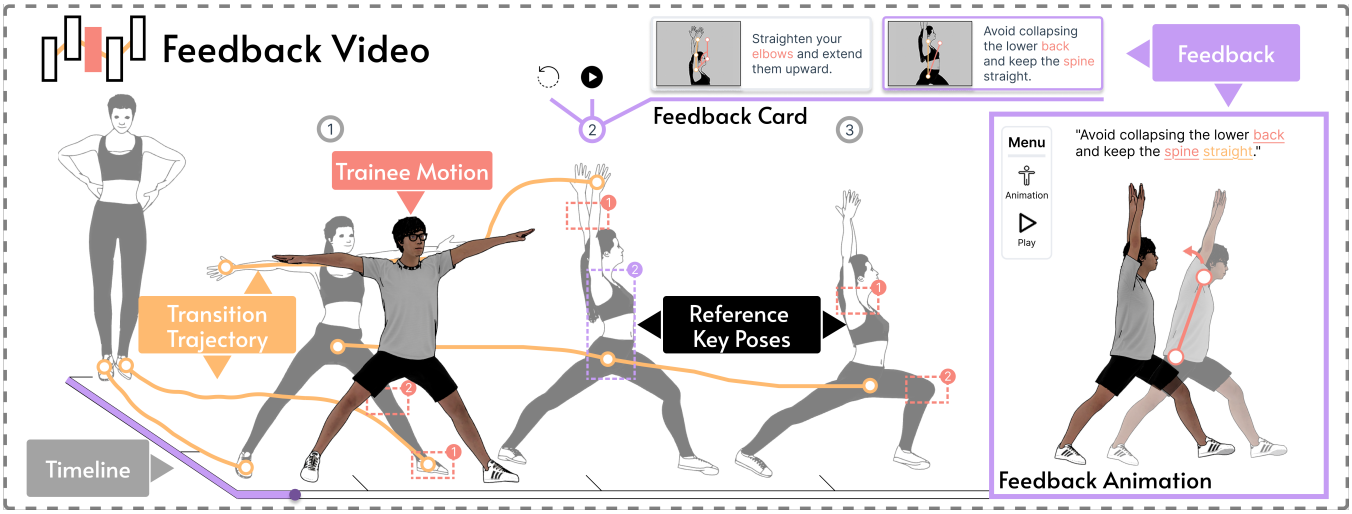

Figure 5: VisMimic generates feedback videos using an integrated motion chain representation. It includes trainee motion (orange), reference key poses (black), transition trajectories (yellow), and supportive UI elements (grey). Dotted orange boxes highlight incorrect body parts, with the numbers indicating the amount of corrective feedback associated with each. Purple highlights denote the currently focused feedback. The design is illustrated using the Yoga Warrior I pose.

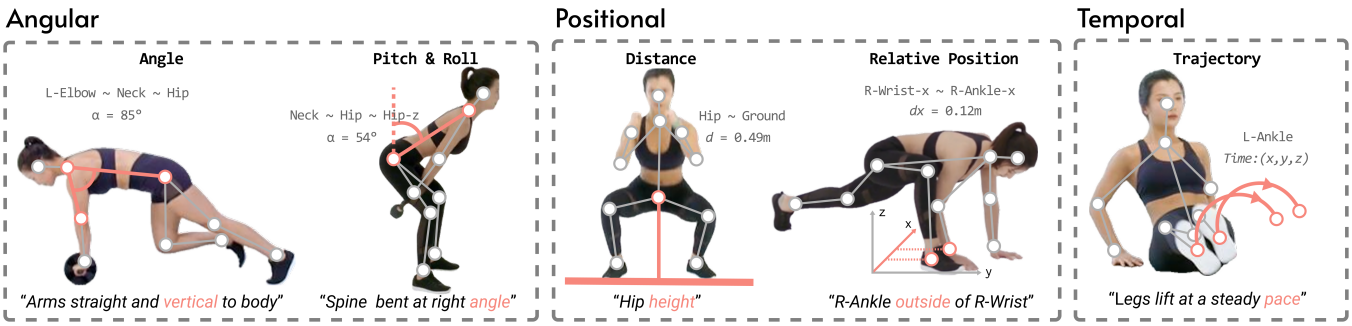

Figure 6: Fine-grained classification of kinematic constraints applied in motion chain. Each illustrates with an example.

Positional. Positional constraints represent the distance and relative position of two key joints. These constraints are defined by a pair of endpoints or their projections on a specific axis. Additionally, the ground is treated as a special point to calculate the height.

- **Distance** describes the L2-distance between two key joints (e.g., the distance between the hip and the ground defines hip height, or the distance between the left and right ankles defines feet distance).
- **Relative Position** describes the difference between the projections of two key points along a given axis, defined by the two key joints and the axis (e.g., the projections of the right wrist and right ankle on the x-axis define which is more inside).

Temporal. The aforementioned attributes describe constraints for fixed static poses. In contrast, temporal constraints represent dynamic changes (position, direction, velocity) in transitional poses. These constraints are defined by sets of key joints. Additionally, velocity is conveyed through the direction and speed of body part movement along the trajectory during dynamic playback, with color encoding reserved for future design extensions [21].

- **Trajectory** describes the spatio-temporal position of joints, encompassing velocity and direction (e.g., the lifting trajectory of the feet can be defined by tracking the ankles).

5.3 VisMimic Workflow

To support create feedback video from two input videos (trainee and reference), we propose a human-AI collaborative workflow comprising four components: motion reconstruction, motion chain definition, feedback generation, and output editor. VisMimic reconstructs 3D human avatars and extracts SMPL parameters [49] from paired videos. The motion chain definition module then extracts key pose frames (e.g., Bottom Phase: the lowest hip position in a squat) and supports setting kinematic constraints. In feedback generation, we involved constraints filter in PoseFix model [24] to analyze key pose pairs and detect motion discrepancies. We use motion principles (e.g., "Keep your upper body as straight as possible" during squat ascent) as external text and cross-reference them with discrepancies to generate the target feedback text. Additionally, we use feedback as motion edit texts to animate the correction of the error pose. Finally, the output editor show the integrated motion

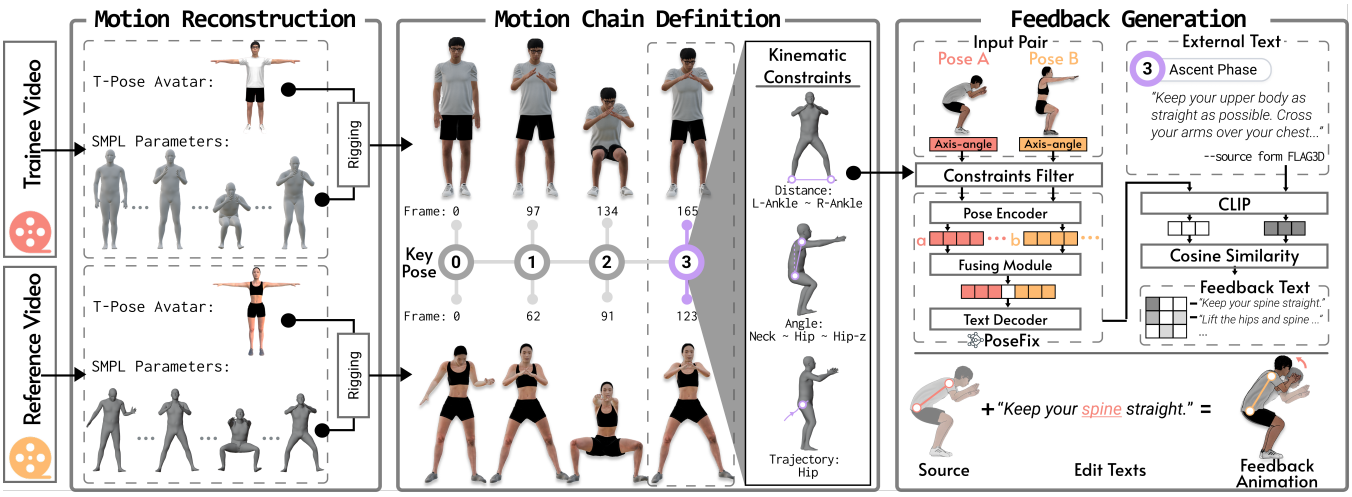

Figure 7: VisMimic workflow encompass motion reconstruction, motion chain definition and feedback generation. It takes a trainee video and a reference video as input, integrates a motion chain structure, and outputs a feedback video. Motion chains serve as data handles, supporting coaches in incorporating their experiential knowledge into the feedback process.

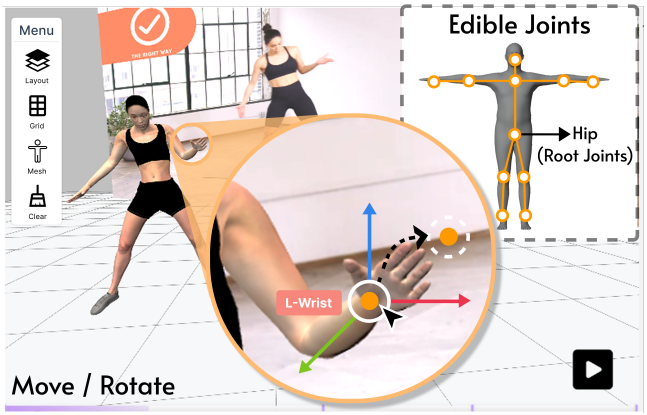

Figure 8: VisMimic overlays the input video as background and supports manual editing of the reconstructed results.

chain representation with feedback, allowing coaches to customize the layout and visual design of the video.

6 Method

In this section, we developed VisMimic, which analyzes trainee videos against reference movements to generate feedback videos.

6.1 Motion Reconstruction

Extracting motion data from video is a common kinematic data source in sports biomechanics for analyzing movement [21, 45]. Using only 2D motion data lacks depth information. While 3D SMPL meshes can simulate muscle deformations during movement [49], the absence of texture details (e.g., clothing and hair) reduces realism. To enable multi-view and high-fidelity animation observation, we leverage state-of-the-art computer vision models [63] to reconstruct 3D motions from input videos (uniformly resampled to 30fps in preprocessing). Drawing on ExAvatar [51], which combines 3D Gaussians [41] and SMPL-X-based surface mesh representations [53], we first generate a full-body 3D avatar from a monocular fitness video in a fixed T-pose. The avatar is then animated with extracted SMPL-X motion parameters and rendered in screen space using 3DGS. Using this method, we generate both the trainee and reference avatars along with their motions.

Validation. While the extracted SMPL-X skeletal and joint motion parameters achieve human-acceptable accuracy, challenges remain due to low-quality or unconstrained in-the-wild video inputs. These can cause motion reconstruction failures, such as physiologically implausible poses (e.g., bent legs in a standing posture) or incorrect joint rotations leading to unnatural movements. To validate reconstruction quality, we provide a display view for visualizing results in SMPL-X or avatar mesh and comparing it with the input video as background [47]. Similar to other animation editing user interface [2], we implement a full-body IK approach with 11 joints (wrists, ankles, elbows, knees, head, neck, and hip), enabling direct manipulation through joint selection and dragging.

6.2 Motion Chain Definition

Based on the motion chain definition and communication with coaches, we first apply it to the FLAG3D dataset [66], a 3D fitness activity dataset with detailed language instructions. Through data cleaning and alignment, we manually structure it into the motion chain format. Then, we developed a semi-automatic processing method to convert reconstructed results—original video (Ori.) and reference video (Ref.)—into a motion chain form, which structures key poses and their relationships. This process includes key pose extraction for motion breakdown and kinematic constraint setting.

6.2.1 Dataset Preparation. We use the FLAG3D dataset [66], which includes 180K videos spanning 60 daily fitness activities, comprising both raw real-world videos and motion capture data. Each activity is categorized by 10 body parts (chest, back, shoulder, arm, neck, abdomen, waist, hip, leg, whole body) and accompanied by detailed, professional sentence-level instructions.

Since fitness activities targeting the same body part share similarities, we select one representative movement from each category.

Working closely with four domain experts (E1-4), we manually process the data in three steps: 1) Identify key poses based on motion capture data and videos. For each motion, we determine key attributes (e.g., hip trajectory in a sumo squat) and visualize them to detect extreme points as key poses, followed by the first round expert verification. 2) Align key pose segmentation with language instructions. Missing instructions are supplemented through expert consensus. 3) Extract initial kinematic constraints from the instructions for each key pose. After ensuring that all experts have a correct understanding of kinematic constraints (Sec. 5.2.2), they independently review and define constraints based on the instructions. Each expert annotated five fitness activities categorized by body part and verified another five annotated by peers. A cross-check process ensured accuracy and completeness, with disagreements resolved through open-meeting discussions following our coding book (see supplementary materials for dataset preparation details).

6.2.2 Key Pose Extraction. Motion-captured data inherently lacks keyframe support, which is widely used in character animation to highlight important moments with representative poses. Prior work [15] detects extreme points in joint trajectories to extract key poses. Inspired by TimeTunnel [84], which extracts key poses from multiple trajectories, we apply a similar approach. We compute the difference between whole-body joint trajectories (11 editable joints, Sec. 6.1) and their Gaussian-smoothed versions frame by frame. Local extrema are identified as key poses, and we iteratively extract a sequence (denoted as A) until the maximum difference falls below a predefined threshold. The threshold is a parameter to constrain the maximum difference in joint distances.

Validation. We evaluate this method on our dataset (30 FPS) under different thresholds. Following action spotting studies [33], we consider a match if the detected key pose aligns with the ground truth (denoted as B) within ±5 frames. Precision is defined as the proportion of elements in A that match B, while recall measures the proportion of B found in A. At a threshold of 0.46, the F1-score peaks at 0.67, providing a reliable cold start for key pose segmentation. Pose extraction, as the first step, demands high accuracy. However, achieving high accuracy remains challenging, as there is a lack of general models that meet coach-acceptable standards for unified pose extraction. Additionally, key poses often carry semantic meaning (e.g., squat, push-up, or tuck jump in a burpee) and may not always align with extreme points. Therefore, for similar fitness activities targeting the same body part (e.g., chest fly and push-ups for the chest), we utilize predefined key attribute templates (e.g., the trajectory and angle of elbow) from our dataset to assist users in refining key poses manually in the display view, enabling them to add, remark, or delete key pose frames as needed.

6.2.3 Kinematic Constraint Setting. After key pose extraction, VisMimic allows users to set the kinematic constraints between key poses with intuitive interaction. As shown in Figure 9, the three types of constraints which we introduced in subsubsection 5.2.2 can be easily defined with clicks. For temporal constraints, i.e., the trajectories of the joints, users can directly click on the joint they want to track. For positional constraints, i.e., the distance constraints, users can first click on a joint or an axis of a joint, and then click on another to constrain the distance between them. For angular constraints, users can click on three objects in order as the first side endpoint, the vertex, and the second side endpoint to define

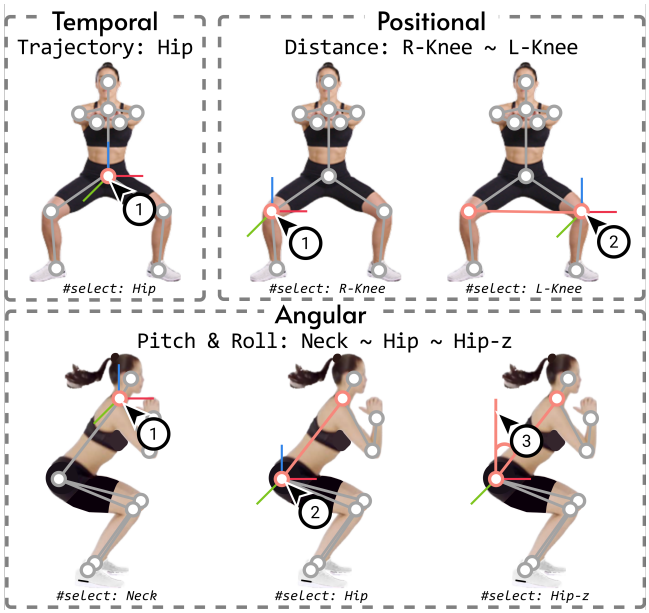

Figure 9: VisMimic supports kinematic constraint setting by allowing users to directly click and link joints and axes.

the angle constraint. The vertex should be a joint, while the two endpoints should be both joints or a joint and an axis of a joint.

6.3 Feedback Generation

Target feedback is generated by comparing key pose pairs from the trainee and reference standard. While few studies focus on providing natural language feedback [20, 29], text alone or with simple key joints markers is often insufficient, placing the burden of how to correct on the trainee [14]. Thus, we first generate the feedback text and then animate it starting from the error pose.

6.3.1 Feedback Text. We aim to generate correctional feedback in natural language, explaining how the trainee's source key pose A should be adjusted to match the target reference pose B. Existing models treat *"human motion as a foreign language"*, mapping motion to natural language for general understanding [37]. For instance, in a squat, the primary focus should be on lower-body movement, while variations such as placing hands behind the head or holding them in front (e.g., sumo squat) are acceptable. However, existing models, lacking semantic awareness, often misidentify upper-body variations as outliers. To enhance these models for feedback generation, we made the following efforts:

1) We first convert our reconstructed SMPL-X parameters (Pose A and Pose B) into joint rotation representations using the SMPL-H model [57] (neglecting facial expressions), employing the axis-angle representation. The global root rotation is then normalized.

2) We adopt PoseFix [24], an auto-regressive model conditioned on pose pairs, as our foundational model. This approach has been applied to generating corrective text for pose modification. However, since our input pairs are key poses with semantic and sequential relationships, directly applying PoseFix results in low-quality feedback. To improve inference, we leverage the hierarchical joint structure of the axis-angle representation to compute global joint positions. Kinematic constraints are incorporated by assigning weighted

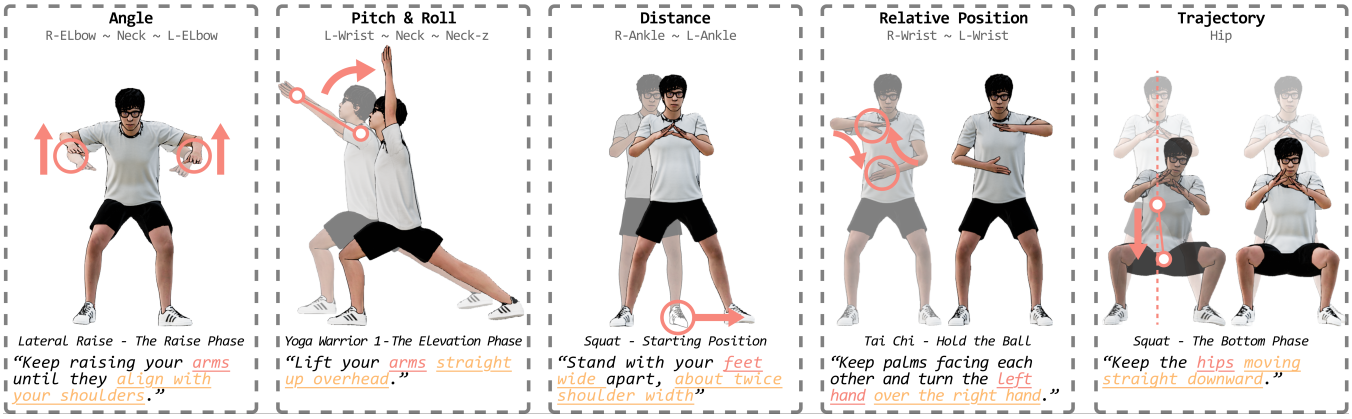

Figure 10: Feedback Gallery: Each kinematic constant includes an example with generated feedback and correctional animation.

importance to related joints (initially $w = 1$). The influence of each joint is propagated and decayed along the parent-child hierarchy. The final normalized weight matrix is applied to the axis-angle pose representation through a constraint filter module. For example, in the bottom phase of a squat, both Pose A and Pose B's axis-angle representation ignore elbow and wrist joints.

3) Pose A and Pose B are encoded as Tokens a and Tokens b, each with a feature dimension of $d = 32$, using a shared pose encoder followed by VPoser architecture [53], modified to support 52 joints of the SMPL-H body model [57]. The encoded tokens are then fused into a set of pose tokens for further processing.

4) The text decoder outputs a probability distribution over the vocabulary for each token. During inference, feedback is generated iteratively using a greedy decoding approach [24].

Validation. To validate the effectiveness of the constraint filter module during inference, we selected three fitness activities from our dataset: sumo squat, push-ups, and jumping jacks. These activities represent lower-body, upper-body, and full-body movements, respectively. For each activity, we annotated key pose errors in 30 samples, with ground truth feedback provided by coaches. We segmented the generated text and ground truth into individual sentences, pairing them for comparison. We then computed accuracy, recall, and Intersection over Union (IoU) metrics. A feedback sentence pair was considered a match if both body part references and descriptive attributes were identical.

Table 1: Feedback text generation results for various fitness activity and with/without constraints filter.

	With Constraints			Without Constraints		
	Acc.	Rec.	IoU	Acc.	Rec.	IoU
Sumo Squat	0.77	0.57	0.49	0.57	0.59	0.41
Push-ups	0.68	0.55	0.44	0.53	0.63	0.40
Jumping Jacks	0.82	0.57	0.51	0.74	0.57	0.47

Recommendation Ranking. For key pose pairs with multiple feedback options, we compare them with existing sentence-level language instructions in our dataset. Each instruction is encoded into tokens using CLIP [56], and their semantic similarity is measured via cosine similarity in the vector space. Feedback is then ranked based on the similarity matrix, ensuring that the most relevant correctional feedback are prioritized. We enable coaches to select feedback in the Correction View (Fig. 11.D), allowing multi-angle comparison of the trainee and reference states in the Compare View (Fig. 11.C1). Relevant kinematic constraints and their numerical values are highlighted to support detailed analysis (Fig. 11.C2). As coaches make selections and additions, feedback for the entire motion chain is progressively generated.

6.3.2 Feedback Animation. To enhance the clarity of generated correctional feedback, we aim to animate the feedback within the motion. Inspired by text-driven motion generation and editing [37], we apply the TMED model from MotionFix [41], a diffusion-based motion editing model. Given the static key pose at the error moment, a textual description detailing the key pose and correctional feedback, and a fixed noise vector to introduce controlled randomness, the model generates an animated correctional motion. The animation naturally transitions toward the corresponding reference key pose as its target. We provide examples, as shown in Fig. 10.

Validation. We tested our method on the previously identified key pose pairs. Using annotated ground truth feedback, we asked 4 coaches to evaluate the motion quality (Mean=3.92, SD=0.64) and plausibility (Mean=4.17, SD=0.55) of the generated animations on a 5-point Likert scale. The results indicate that our approach produces acceptable feedback animations. Additionally, we conducted a quantitative analysis. Since the reference key pose serves as the target pose for correction, we compared the end pose of the generated animation with the corresponding reference key pose. Given that both poses use the same human model with a shared root joint origin, we directly computed the L2 distance using:

$$D(P_1, P_2) = \sqrt{\sum_{i=1}^{n} \|P_{1i} - P_{2i}\|^2}$$

where P_1 and P_2 represent the generated and reference key poses, respectively, and n is the number of joints. Results on all sample pairs (Mean=0.15, SD=0.08) indicate that the feedback can be effectively animated from the source error to the correct target.

Since sequential snapshots effectively illustrate motion patterns, we visualize each feedback's relative constraints from an suggested viewpoint (e.g., front, side, or top). The viewpoint selection depends on the camera angle and the involved body parts. To highlight key

joint movements in the correctional animation, we augment the error pose (Fig. 11.D1) by marking key joints and overlaying directional arrows indicating the movement direction and magnitude.

6.4 Hardware and Software

The models [24, 41, 51] within VisMimic are implemented in PyTorch on a server equipped with an A100 GPU (80GB RAM). The browser interface is built with Vue.js and FastAPI, leveraging HTML Canvas and Three.js for video rendering. By employing WebGPU [81] and multi-threading, we accelerate heavy rendering tasks.

7 Evaluation

We evaluate VisMimic through (1) a real-world use case, (2) a comparative study of generated vs. original feedback videos from coach and trainee perspectives, and (3) a usability study with coaches.

7.1 Usage Scenario

Here we demonstrate how a sports coach, Lexi, creates a feedback video using VisMimic, based on the trainee and reference videos of the sumo squat motion. As a fundamental exercise for glute training, the squat serves as the basis for many complex motions. In this case, Lexi first imports the videos into VisMimic and selects the corresponding motion type (Fig. 11.A). VisMimic automatically reconstructs both the motion and the avatar, displaying the results in the display view with the input video overlaid as a background for visual comparison (Fig. 8). To extract key poses, VisMimic provides an initial segmentation as a cold start. Lexi accepts the reference motion's segmentation but refines the trainee's key frames based on key attributes (e.g., hip height and knee angle).

After reviewing both videos, Lexi switches to the assessment view (Fig. 11.B) to apply constraints on key poses. In addition to using existing templates, she adds custom constraint attributes. By clicking and linking joints and axes, she sets kinematic constraints (e.g., foot distance, where clicking the left ankle, then the right ankle, and linking them specifies the distance constraint). Lexi then uses the compare view (Fig. 11.C1) to inspect the motions from multiple perspectives. A side panel (Fig. 11.C2) presents the numerical values of kinematic attributes, with color encoding highlighting differences. By iteratively refining constraints, Lexi completes the setup. Then VisMimic generates feedback candidates by comparing the trainee's motion against the reference. When Lexi clicks on a feedback candidate, the compare view highlights the corresponding relative constraints, making the correction rationale explicit. After reviewing the feedback suggestions across all key poses, Lexi finalizes and confirms the feedback to be delivered.

The generated feedback video is presented in the edition view. After previewing the video, Lexi can quickly locate and review the feedback presentation by linking with the correction view (Fig. 11.D). The clip view (Fig. 11.D1) provides detailed control and supports customization. For example, since the preparation phase of the squat requires *"feet twice shoulder-width apart"*, Lexi adjusted the observation perspective of the 3clip from side to front for better clarity (as shown in Fig. 12). Finally, a feedback video is created.

7.2 Comparative study

This study aimed to assess the effectiveness of feedback delivery and the viewing experience of generated feedback videos from both the producer's (coach) and recipient's (trainee) perspectives.

7.2.1 Participants. Although VisMimic targets coaches as its primary users, trainees are the ultimate beneficiaries of the feedback videos and should therefore be considered. We recruited 24 participants: 12 sports coaches (P1–P12; 8 male, 4 female; mean age = 29.6, SD = 3.68) and 12 trainees with prior video learning experience (T1–T12; 6 male, 6 female; mean age = 23.7, SD = 3.52). All coaches had at least two years of motor coaching experience and a minimum of three months video-based coaching experience.

7.2.2 Data. We compared feedback videos created by professional coaches with VisMimic generated versions. To ensure data consistency, both versions were created from similar trainee and reference videos with same motion chain structure. The original feedback videos were created independently by three professional coaches (E1-3, from our preliminary study) after identifying key motion errors and feedback instructions. These videos typically included on-screen text and visual cues (e.g., highlights over body parts) to indicate motion errors, followed by demonstrations of correct movements using the reference video. In parallel, we used VisMimic to automatically generate feedback videos conveying the same instructional content. We prepared three video sets covering different motion types — Sumo Squat, Shoulder Press, and Yoga Warrior I — selected to span various body focus areas (lower body, upper body, and full body) and different levels of motion complexity (ranging from simple workout movements to complex yoga poses).

7.2.3 Procedure. This online study lasted approximately 35 minutes, comprising a 30-minute experiment and a 5-minute semi-structured interview. After obtaining recording consent, we collected participant background information and began the session. Aside from background-specific questions, coaches and trainees followed the same procedure and evaluation criteria. During the experiment, participants viewed feedback videos for both one Yoga (7 minutes) and two Workout (4 minutes) scenarios. To minimize potential bias introduced by different motion types, we randomized the order of the video pairs (original vs. VisMimic-generated) for each participant. After individually viewing both versions, participants were asked to complete following tasks: (1) Identification: locate when the motion error occurs and find the body parts with errors; (2) Comprehension: understand how to perform the right movement and how to correct the identified error. Participants then rated both versions using a 7-point Likert scale on feedback effectiveness and viewing experience, followed by selecting the overall preferred version based on these evaluation criteria. A post-study interview was conducted using the "Think Aloud" protocol.

7.3 Usability Study

The purpose of the study was to evaluate the usability of VisMimic with its target users (coaches). The participants were the coaches who had previously taken part in the comparative study. Each study included a 10-minute introduction to VisMimic, followed by hands-on exploration for participants to understand system functionality. After familiarization, coaches were asked to create a feedback video

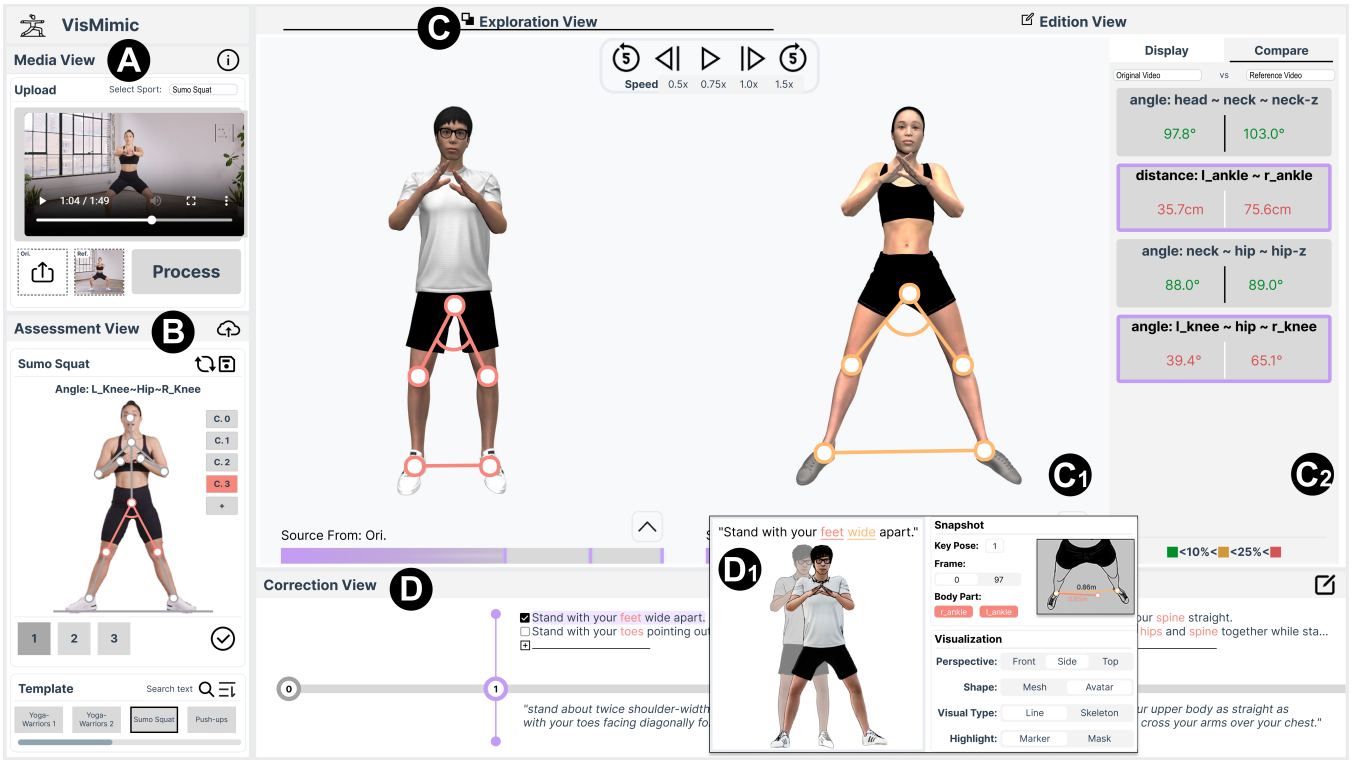

Figure 11: The system interface of VisMimic. (A) The Media view for video management. (B) The assessment view supports interactive kinematic constraint setting. (C) The exploration view displays the reconstructed motion and enables comparison under specified constraints. (D) The correction view presents generated feedback candidates and supports editing.

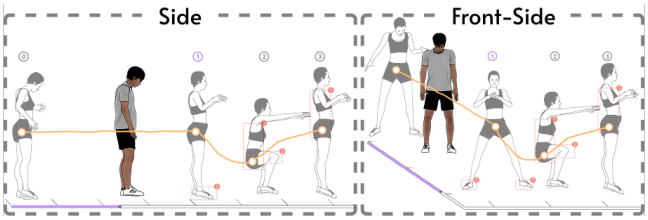

Figure 12: Timeline orientation is consistent with the perspective (side view with X-axis and front view with Y-axis).

within 15 minutes using the same input video pair. During the process, participants were encouraged to share their insights and use VisMimic to edit and customize the content. Afterward, they completed the System Usability Scale (SUS) questionnaire [18]. Finally, we conducted a semi-structured interview to collect user feedback. The entire online session lasted approximately 30 minutes.

7.4 Study Results

7.4.1 Quality of Generated Feedback Video. Based on the 576 ratings from 24 participants, VisMimic-generated feedback videos received favorable scores for feedback delivery effectiveness and viewing experience. Including neutral responses,, 75% (51/72) of user preference ratings favored VisMimic-generated videos over the original version (75% for Yoga case, 50% for Shoulder Press case, and 83% for Sumo Squat case). The differences in participants' evaluations across motion types may be attributed to the varying levels of motion complexity. For simpler motions, such as Shoulder Press, P6 and P8 commented that *"Shoulder Press only contains two key poses, making it easy to learn directly from the motion demonstration"*. In contrast, for more complex motions involving three or more key poses, participants P1, P3, P7 and T4, T5 particularly appreciated the inclusion of key pose snapshots. They noted that these snapshots helped them quickly understand the overall motion pattern and capture essential transitional poses between key movements. We analyzed the questionnaire results as follows:

Effectiveness of feedback delivery. Fig. 13 illustrates that participants generally gave positive ratings for the effectiveness of feedback delivery on both identification and comprehension.

- *Combing key poses with trajectory enables easier identification.* Identifying motion errors in both temporal occurrence and specific body parts is fundamental for motor coaching. Participants provided positive feedback on the identification tasks, including "locating when the error occurs" and "finding the body parts with errors." Specifically, 75.0% of the participants reported that VisMimic facilitated easier error localization. As P2 and T8 noted, *"The red highlight on the key pose directly grabbed my attention, and with subsequent comparisons, I quickly got the motion error."* P5 and P6 appreciated the video navigation enabled by key poses as interaction nodes, commenting that *"these snapshots are indeed representative of the complete motion, making errors easier to notice."* For the error body part identification, 68.1% of the participants found our generated version more effective. Considering that motion errors often involve multiple aspects, both

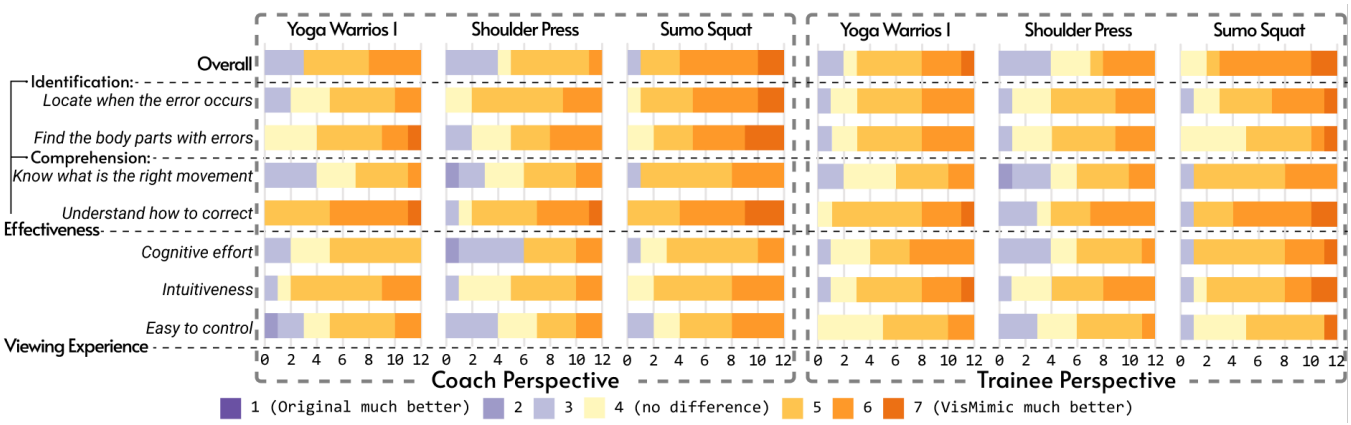

Figure 13: The comparative study evaluates feedback videos from both coach and trainee perspectives, comparing the original version with the VisMimic-generated version.

coaches and trainees found the integration of key pose snapshots particularly helpful for gaining an overview. As P7 reflected, *"The key pose snapshot gave me an overview, and I could further explore details within it, such as trajectory and angles."* Overall, combining key poses with trajectories in feedback video improved users' efficiency in learning differences from the reference standard and supported quickly identifying motion errors.

- *Feedback with correctional animation facilitates easier comprehension.* Enhancing the comprehension of feedback is the primary goal of creating feedback videos in motor skill learning. Participants provided positive feedback on the comprehension-related tasks, including "understanding how to perform the right movement" and "understanding how to correct the error". 63% of the participants agreed that the VisMimic-generated version was more effective in understanding the correct movement execution. As T1 and P10 reported, *"With the video play control and view-switching for front, side, or other perspectives, I could observe the whole motion."* However, for simple motions like Shoulder Press, P5 commented that *"the original video is sufficient as the motion only requires a front view to understand."* Understanding how to correct motion errors was considered a key strength of VisMimic, with nearly 88.9% of participants agreeing on its advantage. Experienced coaches (P1, P6) with over three years of remote coaching experience highly valued the correctional animation, noting that *"previously we told trainees what to do, but it was difficult to describe how to perform it — this visual feedback makes the guidance concrete."* Similarly, T9 and T11 compared it to offline coaching, stating that *"physical contact is effective but limited in remote settings — this animation facilitates communication in remote coaching."* Overall, the correctional animations made feedback more intuitive and easier to comprehend.

Viewing Experience. The overall experience of viewing the feedback videos was well-received. Most participants found the VisMimic-generated version to be more novel and intuitive compared to the original videos. T2 and P5 appreciated the feedback card with clearly labeled error body parts, noting that *"this clearly helped me break down and understand the motion error"*. However, differing from their conventional feedback video experience from both coaches and trainees, the unfamiliar presentation style of the

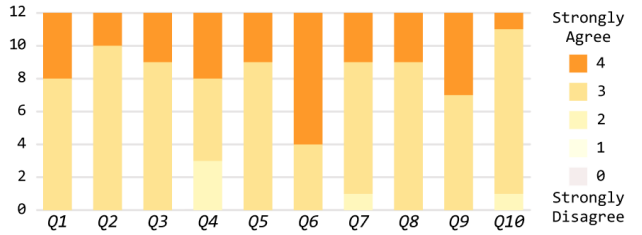

Figure 14: System Usability Scale results of usability study.

VisMimic-generated videos also introduced a certain level of cognitive load. After watching all generated cases, T6, T10, and P11 commented that *"At first, it felt a bit strange, but I gradually got used to this video style and was able to understand the feedback within it"*. In contrast to the original video interaction, P6 and P12 particularly appreciated the multi-view observation and key pose-based navigation provided by VisMimic. They noted that *"It is easy to control the playback to quickly view coaching content"*. T1 and T5 emphasized that *"multi-view is essential for yoga coaching, since occlusion frequently occurs during yoga movements"*. Additionally, P8 and P9 recognized that the motion representation style in our generated video was related to the motion chain. They found that *"this chain with nodes and links is helpful for step-by-step learning."*

7.4.2 Coach Perspective vs. Trainee Perspective. We compared the mean and standard deviation of ratings from coaches and trainees and used pairwise t-tests to assess statistical significance. Results (all $p > 0.05$) indicate no significant differences between the two groups, despite differences in fitness level and understanding of standard movements. This aligns with previous findings and partially supports the overall quality of the generated feedback videos. However, some trainees (T3, T4, and T12) reported that, "the current videos still require considerable cognitive effort, making it difficult to follow and replicate the correct movements." Additionally, age differences may have influenced user experience (trainees on average six years younger than coaches), found the video easier to control. As T7 noted, "The controls are similar to streaming platforms. Despite the amount of information, replaying helps me grasp it easily." Therefore, improving interface accessibility and offering usage guidance could help bridge this gap and enhance the overall user experience across different perspective.

7.4.3 Usability of VisMimic. VisMimic achieved an average SUS score of 81.46, with a learnability score of 76.04 and a usability score of 82.81 (Fig. 14), placing our system within the top 10% of SUS ratings [18], indicating acceptable system usability. Considering that some coaches had limited prior experience with motion analysis systems or primarily relied on just video not motion tracking data, it is reasonable that certain interactions required additional guidance. Despite this, all participants agreed that the system's features worked seamlessly together, and expressed willingness to use VisMimic. This further supports the effectiveness of our interaction design and the integration of modules within the workflow.

8 Discussion

8.1 Reflection on Current Workflow

Applying human biomechanical constraints to improve motion reconstruction results. Most human motion reconstruction methods rely on parametric models such as SMPL [49] and SMPL-X [53], which offer compact, generalizable representations of body shape and pose. These models are widely adopted for their efficiency and compatibility with video-based motion capture pipelines. However, due to the absence of biomechanical constraints, they often generate anatomically or physically implausible poses [76] limiting their accuracy and realism. Incorporating biomechanical constraints [40] into the reconstruction process can reduce such artifacts and enhance the reliability of motion-based feedback.

Incorporating human-object interaction and sports performance in feedback generation. The current version of VisMimic models only the human body, without considering interactions with external objects. However, many motor tasks, such as throwing a ball or swinging a racket, involve human-object interactions where object-related constraints are crucial. Accounting for these interactions would enable more accurate modeling of motion dynamics and allow for more context-aware feedback [14]. Additionally, integrating sports performance metrics, such as output velocity or accuracy [45], can provide meaningful constraints to guide feedback. For example, in a throwing task, the throw speed and precision can serve as measurable outcomes. Incorporating these elements would improve feedback specificity and broaden application scenarios.

Extending VisMimic to Real-Time AI Coaching Applications. Although VisMimic currently functions offline, it holds strong potential for real-time AI coaching deployment. With advancements in video-based motion capture and efficient feedback algorithms, AI analysis and feedback delivery during training sessions become feasible, offering actionable guidance to trainees. To support this interactive paradigm, VisMimic defines a structured input-output pipeline for feedback video generation, enables controllable editing, and incorporates coaches' experiential knowledge via motion chains. Future work will focus on system optimization and model refinement to realize end-to-end AI coaching.

8.2 Limitations and Future Work

We extend the discussion of the current limitations of VisMimic's feasibility and propose feasible opportunities for future work.

- **Feasibility of VisMimic.**
  Applicability: Although VisMimic can process indoor or outdoor motor coaching, its application in the real world may be limited by video quality. High-fidelity motion tracking data are derived from sensors [62] or volumetric video [71]. Our future work will involve incorporating multimodal input into the workflow.
  Generalizability: With model validation, VisMimic achieved acceptable results. However, for in-the-wild scenarios or unique motor tasks (e.g., choreographed movements), the generated feedback may fail (unnatural movements). A feasible solution is to expand the database to include more diverse trainee-reference motion pairs with language feedback [29], covering a wider range of domains. Alternatively, data programming could be employed to define labeling functions that express weak supervision strategies [32], thereby generating large-scale datasets.
  Scalability: (1) With the motion data flow, the system modules are scalable, enabling potential extensions such as interactive editing (e.g., engaging drag operations for motion animation generation [74]). (2) VisMimic currently provides reference-based feedback using normalized motion data. However, this coarse retargeting overlooks individual body differences (e.g., arm length affects movement range), making a universal "standard" difficult to define. In sports like basketball or tennis, such variability further complicates reference selection. To address this, future work will explore physically simulated motor skill experiments [70, 82] to enable more personalized and adaptable feedback.
- **Manual Pose Extraction and Kinematic Constraint Definition.** (1) Accurate pose extraction is critical but challenging, especially for detecting transition poses and handling motion-specific key pose variations. While we tested learning-based methods (e.g., TAL-Net [19], VideoMAE [68]), no general model currently achieves coach-acceptable accuracy, necessitating manual refinement. (2) In fitness coaching, we applied constraint templates for similar motions. However, manually defining kinematic constraints for each key pose is labor-intensive. To reduce this burden, we plan to explore domain-knowledge agents with human motion understanding [73] to automatically extract pose constraints under task-specific semantics.
- **Lack of audio feedback.** VisMimic currently focuses on visual feedback, but audio plays a key role in conveying rhythm and timing in motor coaching. Future work will explore integrating audio to enhance feedback effectiveness.

8.3 Lessons Learned

Implications for video-based coaching. Our user study highlights that coaches place high value on the quality and clarity of feedback outputs. Additionally, there is strong interest in autonomous tools that can streamline the video creation process and reduce manual effort. Coaches expressed a clear need for systems that are not only technically robust but also accessible and time-efficient.

Working with sports experts. Collaborating with sports experts highlighted challenges in adopting new technologies. Many coaches lack familiarity with novel tools. This cognitive gap, which is shaped by diverse backgrounds and varying digital literacy, hinders system adoption. Tasks such as setting constraints or configuring feedback require intuitive interfaces and clear guidance. Building mutual understanding through close collaboration was essential to ensure usability and trust in the system.

Multi-perspective evaluation reflects the dual nature of coaching. The coaching process is inherently bidirectional, where feedback videos serve as a communication medium which are created by coaches and interpreted by trainees. As a generation tool,

VisMimic's output quality must address the needs of both. Including both perspectives in evaluation ensures that videos are not only easy to produce but also effective and actionable for the end users.

9 Conclusion

We introduced VisMimic, a novel approach to assist coaches in creating augmented feedback videos for motor skill learning through the comparison of trainee and reference videos. For feedback representation, we integrated the motion chain into feedback videos, combining key poses with motion trajectories to simultaneously convey both the overall motion dynamics and critical static postures. For feedback generation, our method automatically generates feedback candidates and delivers feedback in both textual descriptions and visual animation. User studies demonstrate the effectiveness of the generated feedback videos in enhancing error identification and movement comprehension, and the usability of VisMimic.

Acknowledgments

The work was supported by NSFC (U22A2032, 62421003, 62402437), and the Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government (ZJU).

References

[1] 2025. AasanAI. https://github.com/saRvaGnyA/AasanAI
[2] 2025. Blender. http://www.blender.org
[3] 2025. BodyPark. https://bodypark.cn/home
[4] 2025. Keep. https://keep.com/
[5] 2025. MotionPro. https://www.motionprosoftware.com
[6] 2025. Myogai. https://www.myogai.com/
[7] 2025. Sportsbox AI. https://www.sportsbox.ai/
[8] 2025. SwingVision. https://swing.vision/
[9] 2025. VisualEyes. https://www.visualeyesapp.com/
[10] 2025. Volt Athletics. https://voltathletics.com/
[11] 2025. YogiFi | Your Smart Yoga Mat. https://yogifismart.com/
[12] Omid Alemi, Philippe Pasquier, and Chris D. Shaw. 2014. Mova: Interactive Movement Analytics Platform. *Proceedings of the 2014 International Workshop on Movement and Computing* (2014), 37–42. https://doi.org/10.1145/2617995.2618002
[13] Peter O'Donoghue and. 2006. The use of feedback videos in sport. *International Journal of Performance Analysis in Sport* 6, 2 (2006), 1–14. https://doi.org/10.1080/24748668.2006.11868368
[14] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. 2024. ExpertAF: Expert Actionable Feedback from Video. *ArXiv* abs/2408.00672 (2024). https://doi.org/10.48550/ARXIV.2408.00672
[15] Jackie Assa, Yaron Caspi, and Daniel Cohen-Or. 2005. Action synopsis: pose selection and illustration. *ACM Trans. Graph.* 24, 3 (2005), 667–676. https://doi.org/10.1145/1073204.1073246
[16] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. 2024. MotionFix: Text-Driven 3D Human Motion Editing. In *SIGGRAPH Asia 2024 Conference Papers*. https://doi.org/10.1145/3680528.3687559
[17] Jorik Blaas, Charl Botha, Edward Grundy, Mark Jones, Robert Laramee, and Frits Post. 2009. Smooth Graphs for Visual Exploration of Higher-Order State Transitions. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 969–976. https://doi.org/10.1109/TVCG.2009.181
[18] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[19] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the Faster R-CNN Architecture for Temporal Action Localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1130–1139. https://doi.org/10.1109/CVPR.2018.00124
[20] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. 2024. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. *ArXiv* abs/2405.20340 (2024).
[21] Liqi Cheng, Hanze Jia, Lingyun Yu, Yihong Wu, Shuainan Ye, Dazhen Deng, Hui Zhang, Xiao Xie, and Yingcai Wu. 2024. VisCourt: In-Situ Guidance for Interactive Tactic Training in Mixed Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3654777.3676466
[22] Christopher Clarke, Doga Cavdir, Patrick Chiu, Laurent Denoue, and Don Kimber. 2020. Reactive Video: Adaptive Video Playback Based on User Motion for Supporting Physical Activity. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 196–208. https://doi.org/10.1145/3379337.3415591
[23] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 2022. PoseScript: 3D Human Poses from Natural Language. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. 346–362. https://doi.org/10.1007/978-3-031-20068-7_20
[24] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. 2023. PoseFix: Correcting 3D Human Poses with Natural Language. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14972–14982. https://doi.org/10.1109/ICCV51070.2023.01379
[25] Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N. Balasubramanian, Bharathi Callepalli, and Ayon Sharma. 2022. Pose Tutor: An Explainable System for Pose Correction in the Wild. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 3539–3548. https://doi.org/10.1109/CVPRW56347.2022.00398
[26] Bhat Dittakavi, Bharathi Callepalli, Aleti Vardhan, Sai Vikas Desai, and Vineeth N Balasubramanian. 2024. CARE: Counterfactual-based Algorithmic Recourse for Explainable Pose Correction. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4890–4899. https://doi.org/10.1109/WACV57701.2024.00483
[27] Linfeng Dong, Wei Wang, Yu Qiao, and Xiao Sun. 2024. LucidAction: A Hierarchical and Multi-model Dataset for Comprehensive Action Quality Assessment. In *Neural Information Processing Systems*.
[28] Augusto Dias Pereira dos Santos, Lian Loke, and Roberto Martínez Maldonado. 2018. Exploring video annotation as a tool to support dance teaching. *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (2018). https://doi.org/10.1145/3292147.3292194
[29] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. 2021. AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9914–9923. https://doi.org/10.1109/CVPR46437.2021.00979
[30] C. Gosselin and J. Angeles. 1990. Singularity analysis of closed-loop kinematic chains. *IEEE Transactions on Robotics and Automation* 6, 3 (1990), 281–290. https://doi.org/10.1109/70.56660
[31] Aditya Gunturu, Yi Wen, Nandi Zhang, Jarin Thundathil, Rubaiat Habib Kazi, and Ryo Suzuki. 2024. Augmented Physics: Creating Interactive and Embedded Physics Simulations from Static Textbook Diagrams. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3654777.3676392
[32] Yuchen He, Jianbing Lv, Liqi Cheng, Lingyu Meng, Dazhen Deng, and Yingcai Wu. 2025. ProTAL: A Drag-and-Link Video Programming Framework for Temporal Action Localization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18. https://doi.org/10.1145/3706598.3713741
[33] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 961–970. https://doi.org/10.1109/CVPR.2015.7298698
[34] Yueqi Hu, Shuangyuan Wu, Shi hong Xia, Jinghua Fu, and Wei Chen. 2010. Motion track: Visualizing variations of human motion data. *2010 IEEE Pacific Visualization Symposium (PacificVis)* (2010), 153–160. https://doi.org/10.1109/PACIFICVIS.2010.5429596
[35] Keiichi Ihara, Kyzyl Monteiro, Mehrad Faridan, Rubaiat Habib Kazi, and Ryo Suzuki. 2025. Video2MR: Automatically Generating Mixed Reality 3D Instructions by Augmenting Extracted Motion from 2D Videos. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1548–1563. https://doi.org/10.1145/3708359.3712159
[36] Sujin Jang, Niklas Elmqvist, and Karthik Ramani. 2016. MotionFlow: Visual Abstraction and Aggregation of Sequential Patterns in Human Motion Tracking Data. *IEEE Transactions on Visualization and Computer Graphics* 22 (2016), 21–30. https://doi.org/10.1109/TVCG.2015.2468292
[37] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. In *Advances in Neural Information Processing Systems*, Vol. 36. 20067–20079.
[38] Hye-Young Jo, Laurenz Seidel, Michel Pahud, Mike Sinclair, and Andrea Bianchi. 2023. FlowAR: How Different Augmented Reality Visualizations of Online Fitness Videos Support Flow for At-Home Yoga Exercises. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580897
[39] Raine Kajastila, Leo Holsti, and Perttu Hämäläinen. 2016. The Augmented Climbing Wall: High-Exertion Proximity Interaction on a Wall-Sized Interactive Surface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 758–769. https://doi.org/10.1145/2858036.2858450
[40] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. 2023. From Skin to Skeleton: Towards Biomechanically

Accurate 3D Digital Humans. *ACM Trans. Graph.* 42, 6 (2023). https://doi.org/10.1145/3618381

[41] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023). https://doi.org/10.1145/3592433

[42] Ang Li, Jiazhou Liu, Maxime Cordeil, Jack Topliss, Thammathip Piumsomboon, and Barrett Ens. 2023. GestureExplorer: Immersive Visualisation and Exploration of Gesture Data. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580678

[43] Lei Li, Sen Jia, Jianhao Wang, Zhaochong An, Jiaang Li, Jenq-Neng Hwang, and Serge Belongie. 2025. ChatMotion: A Multimodal Multi-Agent for Human Motion Analysis. *ArXiv* abs/2502.18180 (2025).

[44] William Li, L. R. Bartram, and Philippe Pasquier. 2016. Techniques and Approaches in Static Visualization of Motion Capture Data. *Proceedings of the 3rd International Symposium on Movement and Computing* (2016). https://doi.org/10.1145/2948910.2948935

[45] Tica Lin, Rishi Singh, Yalong Yang, Carolina Nobre, Johanna Beyer, Maurice A. Smith, and Hanspeter Pfister. 2021. Towards an Understanding of Situated AR Visualization for Basketball Free-Throw Training. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445649

[46] Jingyuan Liu, Nazmus Saquib, Zhutian Chen, Rubaiat Habib Kazi, Li-Yi Wei, Hongbo Fu, and Chiew-Lan Tai. 2022. PoseCoach: A Customizable Analysis and Visualization System for Video-Based Running Coaching. *IEEE Transactions on Visualization and Computer Graphics* 30 (2022), 3180–3195. https://doi.org/10.1109/TVCG.2022.3230855

[47] Jingyuan Liu, Li-Yi Wei, Ariel Shamir, and Takeo Igarashi. 2024. iPose: Interactive Human Pose Reconstruction from Video. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3641944

[48] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. 2024. Bridging the Gap Between Human Motion and Action Semantics via Kinematic Phrases. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*. 223–240. https://doi.org/10.1007/978-3-031-73242-3_13

[49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 16 pages. https://doi.org/10.1145/2816795.2818013

[50] Dizhi Ma, Xiyun Hu, Jingyu Shi, Mayank Patel, Rahul Jain, Ziyi Liu, Zhengzhe Zhu, and Karthik Ramani. 2024. avaTTAR: Table Tennis Stroke Training with On-body and Detached Visualization in Augmented Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3654777.3676400

[51] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. 2024. Expressive Whole-Body 3D Gaussian Avatar. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLI*. 19–35. https://doi.org/10.1007/978-3-031-72940-9_2

[52] Masaki Oshita. 2019. Motion Volume: Visualization of Human Motion Manifolds. *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry* (2019). https://doi.org/10.1145/3359997.3365684

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10967–10977. https://doi.org/10.1109/CVPR.2019.01123

[54] F. Potdevin, O. Vors, A. Huchez, M. Lamour, K. Davids, and C. Schnitzler and. 2018. How can video feedback be used in physical education to support novice learning in gymnastics? Effects on motor learning, self-assessment and motivation. *Physical Education and Sport Pedagogy* 23, 6 (2018), 559–574. https://doi.org/10.1080/17408989.2018.1485138

[55] A. Johannes Pretorius and Jarke J. Van Wijk. 2006. Visual Analysis of Multivariate State Transition Graphs. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 685–692. https://doi.org/10.1109/TVCG.2006.192

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*.

[57] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.* 36, 6 (2017). https://doi.org/10.1145/3130800.3130883

[58] Yasuhiko Sakamoto, Shigeru Kuriyama, and Toyohisa Kaneko. 2004. Motion map: image-based retrieval and segmentation of motion data. In *Symposium on Computer Animation*. 259–266.

[59] Muhammad Atif Sarwar, Yu-Chen Lin, Yousef-Awwad Daraghmi, Tsì-Uí Ík, and Yih-Lang Li. 2023. Skeleton Based Keyframe Detection Framework for Sports Action Analysis: Badminton Smash Case. *IEEE Access* 11 (2023), 90891–90900. https://doi.org/10.1109/ACCESS.2023.3307620

[60] Alessandra Semeraro and Laia Turmo Vidal. 2022. Visualizing Instructions for Physical Training: Exploring Visual Cues to Support Movement Learning from Instructional Videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517735

[61] Ziwei Shan, Yaoyu He, Chengfeng Zhao, Jiashen Du, Jingyan Zhang, Qixuan Zhang, Jingyi Yu, and Lan Xu. 2025. Mojito: LLM-Aided Motion Instructor with Jitter-Reduced Inertial Tokens. *ArXiv* abs/2502.16175 (2025).

[62] Tal Shany, Stephen J. Redmond, Michael R. Narayanan, and Nigel H. Lovell. 2012. Sensors-Based Wearable Systems for Monitoring of Human Movement and Falls. *IEEE Sensors Journal* 12, 3 (2012), 658–670. https://doi.org/10.1109/JSEN.2011.2146246

[63] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia Conference Proceedings*. https://doi.org/10.1145/3680528.3687565

[64] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. 2023. PhaseMP: Robust 3D Pose Estimation via Phase-conditioned Human Motion Prior. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14679–14691. https://doi.org/10.1109/ICCV51070.2023.01353

[65] Richard Tang, Hesam Alizadeh, Anthony Tang, Scott Bateman, and Joaquim A.P. Jorge. 2014. Physio@Home: design explorations to support movement guidance. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. 1651–1656. https://doi.org/10.1145/2559206.2581197

[66] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. 2023. FLAG3D: A 3D Fitness Activity Dataset with Language Instruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22106–22117. https://doi.org/10.1109/CVPR52729.2023.02117

[67] Feng Tian, Shuting Ni, Xiaoyue Zhang, Fei Chen, Qiaolian Zhu, Chunyi Xu, and Yuzhi Li. 2024. Enhancing Tai Chi Training System: Towards Group-Based and Hyper-Realistic Training Experiences. *IEEE Transactions on Visualization and Computer Graphics* 30, 5 (2024), 1–11. https://doi.org/10.1109/TVCG.2024.3372099

[68] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. https://doi.org/10.5555/3600270.3601002

[69] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. 2019. AI Coach: Deep Human Pose Estimation and Analysis for Personalized Athletic Training Assistance. *Proceedings of the 27th ACM International Conference on Multimedia* (2019), 374–382. https://doi.org/10.1145/3343031.3350910

[70] Yinhuai Wang, Qihan Zhao, Runyi Yu, Ailing Zeng, Jing Lin, Zhengyi Luo, Hok Wai Tsui, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Tan Ping. 2024. SkillMimic: Learning Reusable Basketball Skills from Demonstrations. *arXiv preprint arXiv:2408.15270* (2024).

[71] Jiqing Wen, Lauren Gold, Qianyu Ma, and Robert LiKamWa. 2024. Augmented Coach: Volumetric Motion Annotation and Visualization for Immersive Sports Coaching. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 137–146. https://doi.org/10.1109/VR58804.2024.00037

[72] Krist Wongsuphasawat and David Gotz. 2012. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2659–2668. https://doi.org/10.1109/TVCG.2012.225

[73] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2024. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. *arXiv preprint arXiv: 2405.17013*.

[74] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. 2024. DragAnything: Motion Control for Anything Using Entity Representation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXII*. 331–348. https://doi.org/10.1007/978-3-031-72670-5_19

[75] Yihong Wu, Lingyun Yu, Jie Xu, Dazhen Deng, Jiachen Wang, Xiao Xie, Hui Zhang, and Yingcai Wu. 2023. AR-Enhanced Workouts: Exploring Visual Cues for At-Home Workout Videos in AR Environment. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606796

[76] Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. 2025. Reconstructing Humans with a Biomechanically Accurate Skeleton. *ArXiv* (2025).

[77] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. 2024. MotionBank: A Large-scale Video Motion Benchmark with Disentangled Rule-based Annotations. *ArXiv* abs/2410.13790 (2024).

[78] Jingyu Yan and Marc Pollefeys. 2008. A Factorization-Based Approach for Articulated Nonrigid Shape, Motion and Kinematic Chain Recovery From Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 5 (2008), 865–877. https://doi.org/10.1109/TPAMI.2007.70739

[79] Lijie Yao, Anastasia Bezerianos, Romain Vuillemot, and Petra Isenberg. 2022. Visualization in Motion: A Research Agenda and Two Evaluations. *IEEE*

Transactions on Visualization and Computer Graphics 28, 10 (2022), 3546–3562. https://doi.org/10.1109/TVCG.2022.3184993

[80] Hiroshi Yasuda, Ryota Kaihara, Suguru Saito, and Masayuki Nakajima. 2008. Motion Belts: Visualization of Human Motion Data on a Timeline. *IEICE Trans. Inf. Syst.* 91-D (2008), 1159–1167. https://doi.org/10.1093/ietisy/e91-d.4.1159

[81] Geng Yu, Chang Liu, Ting Fang, Jinyuan Jia, Enming Lin, Yiqiang He, Siyuan Fu, Long Wang, Lei Wei, and Qingyu Huang. 2023. A survey of real-time rendering on Web3D application. *Virtual Reality & Intelligent Hardware* 5, 5 (2023), 379–394. https://doi.org/10.1016/j.vrih.2022.04.002

[82] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. 2023. Learning Physically Simulated Tennis Skills from Broadcast Videos. *ACM Trans. Graph.* 42, 4 (2023). https://doi.org/10.1145/3592408

[83] Ziyi Zhao, Sena Kiciroglu, Hugues Vinzant, Yuan-Chen Cheng, Isinsu Katircioglu, Mathieu Salzmann, and P. Fua. 2023. 3D Pose Based Feedback for Physical Exercises. In *Computer Vision – ACCV 2022.*

[84] Qian Zhou, David Ledo, George Fitzmaurice, and Fraser Anderson. 2024. TimeTunnel: Integrating Spatial and Temporal Motion Editing for Character Animation in Virtual Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3613904.3641927

[85] Chen Zhu-Tian, Shuainan Ye, Xiangtong Chu, Haijun Xia, Hui Zhang, Huamin Qu, and Yingcai Wu. 2022. Augmenting Sports Videos with VisCommentator. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 824–834. https://doi.org/10.1109/TVCG.2021.3114806