

# ViSTAR: Virtual Skill Training with Augmented Reality with 3D Avatars and LLM coaching agent

Chunggi Lee\*  
chunggi\_lee@g.harvard.edu  
Harvard University  
Cambridge, Massachusetts, USA

Eiji Ikeda  
iked.eiji.ga@u.tsukuba.ac.jp  
University of Tsukuba  
Tsukuba, Japan

Hayato Saiki\*  
saiki@ai.iit.tsukuba.ac.jp  
University of Tsukuba  
Tsukuba, Japan

Kenji Suzuki  
kenji@ieee.org  
University of Tsukuba  
Tsukuba, Japan

Tica Lin  
mlin@g.harvard.edu  
Dolby Laboratories  
Atlanta, Georgia, USA

Chen Zhu-Tian  
ztchen@umn.edu  
University of Minnesota-Twin Cities  
Minneapolis, Minnesota, USA

Hanspeter Pfister  
pfister@seas.harvard.edu  
Harvard University  
Cambridge, Massachusetts, USA

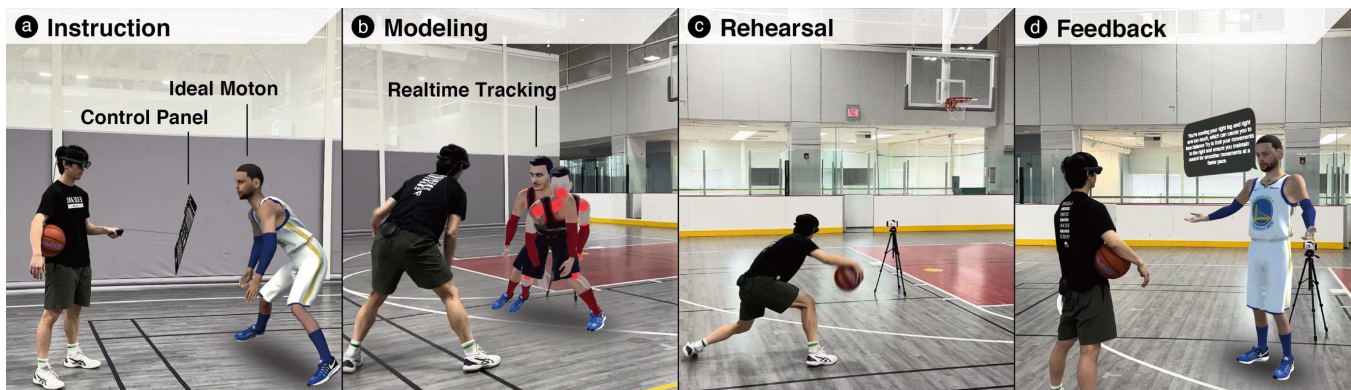


Figure 1: ViSTAR is designed based on the Behavioral Skills Training (BST) teaching framework with four key steps: instruction, modeling, rehearsal, and feedback. Learners can (a) receive instruction of ideal motion demonstrated by a 3D animated avatar, (b) break down the action into key segments with real-time expert overlays, (c) rehearse the movement by recording their own performance, and (d) receive comprehensive multi-faceted feedback from a virtual coach in verbal and visual form.

## Abstract

We present ViSTAR, a Virtual Skill Training system in AR that supports self-guided basketball skill practice, with feedback on balance, posture, and timing. From a formative study with basketball players and coaches, the system addresses three challenges: understanding skills, identifying errors, and correcting mistakes. ViSTAR follows the Behavioral Skills Training (BST) framework—instruction, modeling, rehearsal, and feedback. It provides feedback through visual overlays, rhythm and timing cues, and an AI-powered coaching agent using 3D motion reconstruction. We generate verbal feedback by analyzing spatio-temporal joint data and mapping features to

natural-language coaching cues via a Large Language Model (LLM). A key novelty is this feedback generation: motion features become concise coaching insights. In two studies ( $N=16$ ), participants generally preferred our AI-generated feedback to coach feedback and reported that ViSTAR helped them notice posture and balance issues and refine movements beyond self-observation.

## CCS Concepts

• Human-centered computing → Interactive systems and tools.

## Keywords

Embodied Skill Training, Augmented Reality, Large Language Model

## ACM Reference Format:

Chunggi Lee\*, Hayato Saiki\*, Tica Lin, Eiji Ikeda, Kenji Suzuki, Chen Zhu-Tian, and Hanspeter Pfister. 2026. ViSTAR: Virtual Skill Training with Augmented Reality with 3D Avatars and LLM coaching agent. In *Proceedings*



This work is licensed under a Creative Commons Attribution 4.0 International License. CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2278-3/2026/04  
<https://doi.org/10.1145/3772318.3790634>

of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3772318.3790634>

## 1 INTRODUCTION

Motor skill acquisition is widely described as an embodied process across domains such as dance, music, craft, and sports [4, 5, 41, 49, 83, 84]. In this view, learners gradually align how movements *feel* with how they are actually *performed*, cultivating an internal sense of balance, rhythm, weight transfer, and timing. In competitive sports, these embodied skills directly shape athletes' ability to create scoring opportunities and respond to rapidly changing game situations. Because these skills are grounded in internal sensations rather than explicit rules, athletes often find it difficult to identify and correct their own technique. As a result, motor skill acquisition is considered highly relevant to coaching practice, with significant practical implications for training and performance [26, 81].

Embodied motor learning is fundamental for both recreational and elite athletes [45, 74, 79]. While coach-guided training effectively translates internal sensations into actionable techniques [18, 30], such personalized guidance is resource-intensive and largely inaccessible to non-professional athletes [67]. Timely, precise feedback is critical for refining technique [17, 37, 52, 81], yet remains largely inaccessible for complex and multi-step movements. To support motor learning without a coach, prior research has leveraged head-mounted displays (HMDs)-based immersive training across various sports and dance [14, 17, 24, 28, 37, 52]. While promising for tactical decision-making or repeatable skills (e.g., free-throws), current HMDs introduce practical constraints in high-intensity sports (e.g., fast movements and limited field of view), which limit their ecological feasibility in full gameplay. As a result, complex motor skills that involve multiple phases and coordinated limb actions (e.g., advanced basketball dribbling and footwork) remain relatively underexplored in Augmented Reality (AR)/Virtual Reality (VR) coaching systems.

Motivated by the gap in providing guidance, we investigate an AR- and AI-driven system. This system focuses on kinematics, utilizing external feedback as a scaffold for the athlete's own reflection on their internal bodily awareness, thereby helping them recognize and interpret technical errors. We focus on controlled, demanding basketball moves where HMD use is practical, employing Behavioral Skills Training (BST) [70] as a pedagogical scaffold to sequence multi-phase AR guidance. Our design goal is to use these rich external cues (e.g., joint angles and posture) to help athletes reflect on their felt sense of balance, rhythm, and weight transfer. Critically, ViSTAR provides kinematic feedback and verbal guidance to scaffold this reflection, but we explicitly do not attempt to directly measure or train inner bodily sensations, which remain outside the reach of current non-invasive hardware. Through a user-centered design process, we developed ViSTAR, an AR training system that provides basketball learners with personalized coaching feedback for isolated motor skills using 3D avatars and multi-faceted, LLM-powered feedback. At a high level, ViSTAR adopts the four components of BST—*instruction*, *modeling*, *rehearsal*, and *feedback* [70]—as the backbone of the training workflow. In

the *instruction* and *modeling* phases, ViSTAR uses 3D avatars and segmented exemplars to present a 360° view of the move and support step-by-step visual alignment. During *rehearsal*, learners practice the move while ViSTAR captures their motion, and in the *feedback* phase, we provide side-by-side comparison, hit judgments, and verbal feedback that highlight which parts of the motion need adjustment. In designing feedback, we explore two levels of visual guidance: (1) **Holistic Motion**, which summarizes overall movement quality with a focus on flow and timing, and (2) **Localized Motion**, which targets specific motion segments (Sec. 4). Because verbal feedback is central to effective coaching, we also contribute an LLM-powered pipeline that translates joint-level motion analysis into natural-language coaching cues by (1) transforming motions into textual descriptors, (2) using Random Forest decision paths to identify salient motion differences, and (3) leveraging an LLM to generate actionable, context-aware feedback. To address the lack of datasets containing paired motions of the same skill, we simulate learner executions by injecting joint-level perturbations and use Random Forest to prioritize joint differences, making this data-to-language transformation feasible and scalable.

We conducted two user studies to examine the quality of AI-generated coaching feedback and the perceived usefulness of the system. The first compared participants' preferences between AI and human coach feedback, finding participants preferred AI for identifying posture errors and providing concrete, actionable corrections. The second study compared traditional self-observation with ViSTAR's AR system (integrating a 3D avatar and LLM guidance). While the small sample size limits strong claims about performance improvement, participants reported that ViSTAR helped them understand and recognize errors, rating the AR training as both engaging and useful.

In summary, we make the following contributions: (1) ViSTAR, an AR skill training system leveraging reconstructed 3D avatars and multi-faceted feedback for basketball practice; (2) a feedback design framework combining holistic (flow/rhythm) and localized (joint/path) guidance to visualize posture, timing, and weight transfer; (3) a method for generating AI coaching feedback by linking joint-level motion analysis with Large Language Models; and (4) an empirical evaluation (ViSTAR) through two user studies, providing insights into AI vs. coach feedback and how AR guidance supports self-directed learning. Given the small-sample, short-term constraints, we position ViSTAR as an enabling AR+AI platform that supports reflection on embodied aspects of skill learning, rather than a definitive performance-enhancing tool.

## 2 RELATED WORK

### 2.1 Sports Training in Immersive Systems

Immersive technologies such as Virtual Reality (VR) and Augmented Reality (AR) have been widely explored for motor skill training in sports. These systems are typically categorized into VR-based and AR-based approaches.

**VR-based training systems** have gained significant traction for their high immersion and control, with several studies showing their effectiveness in training tactical and perceptual skills in basketball. For example, Tsai et al. [75] developed a system that

\*These authors contributed equally to this research.

allows players to rehearse basketball tactics from global or player-specific perspectives, with virtual defenders reacting to head pose. In a follow-up study, they [76] used a multi-camera setup and VR device to evaluate performance, finding that increased immersion enhances understanding of offensive strategies and strategic imagery. VisionCoach [56] trains visual scanning and passing decisions through VR-based game scenarios. A number of review studies have thoroughly examined the use of VR in sports training, focusing on performance assessment and training in team ball sports [1], applications for competitive athletes [25], and the integration of interactive VR systems in sport [63]. While promising in controlled environments, VR systems replace the physical training context, which can limit realistic movement, hinder transfer to actual courts, and raise safety concerns. These limitations underscore the value of AR-based systems, which embed guidance and feedback directly into the physical environment, enabling more natural interaction and contextual relevance.

**AR-based training systems**, in contrast, embed digital guidance into the real world, allowing users to stay aware of their surroundings while receiving visual or interactive cues. This context-aware augmentation supports safer, more natural practice, particularly for full-body and sport-specific movements. Several systems have leveraged AR to support physical skill learning. YouMove [2] used a large-scale AR mirror to guide users through recorded movements, fading cues over time to encourage retention. Building on this idea, AR-Enhanced Workouts [85] employed pose-based overlays for home fitness, showing that situated feedback improves understanding and real-time correction. Lin et al. [52] visualized basketball shot trajectories in real time, enhancing shooting consistency and form awareness. Tai Chi AR Trainer [15] and Soccer MR Trainer [42] extended AR guidance to martial arts and outdoor sports, providing pose evaluation via head-mounted displays. In the context of team sports, VisCourt [17] delivered in-situ tactical instruction for basketball, enabling multiplayer coordination and decision-making in real physical environments. Despite their benefits, existing systems often lack personalization and rely on limited binary or part-level feedback, without leveraging AI to interpret nuanced motion data. In contrast, our system is designed to support more dynamic and personalized motor training by adapting to individual performance and delivering fine-grained, context-aware feedback through integrated spatio-temporal analysis and AI-powered interpretation.

## 2.2 Feedback Mechanisms for Embodied Motor Learning

**Visual Feedback for Human Motor Training.** Visualization plays a key role in exploring human motion data, helping researchers understand movement patterns and communicate insights effectively. When integrated into immersive environments, such visualizations enhance observation and deepen understanding of complex motion. A common method is to visualize 3D trajectories of body parts (e.g., hands, head, feet) over time [31, 43]. For instance, MIRIA [10] provides 3D trajectories, heatmaps, and scatterplots from recorded movements and interactions in mixed reality, supporting behavior analysis. However, such visualizations often

emphasize global position and overlook detailed posture, limiting contextual interpretation. To address this, skeleton- or avatar-based visualizations have been adopted to more realistically represent posture and gesture [6, 12, 13, 36, 44], allowing users to observe motion with minimal cognitive load. Wu et al. [85] proposed a design space for visualizing workouts, categorizing visualizations by task, data type, and spatial relation to the body, and offering practical guidance for designing AR-based motion visualizations.

**Embodied Skill Learning Beyond Sports.** From an embodied learning perspective, motor skill acquisition involves more than reducing kinematic error: learners gradually align internal sensations of balance, rhythm, timing, and weight transfer with external cues and instructional structures [4, 8, 41, 49, 84]. HCI and the learning sciences have examined such embodied motor learning in domains beyond sports. VR and AR systems for dance and yoga use motion capture, avatar guidance, and overlaid videos to support timing, rhythm, and expressive movement flow, not only spatial accuracy [13, 39]. In music education, embodied accounts of instrumental learning emphasize how students couple proprioceptive sensations with auditory outcomes and teacher feedback to refine technique [49, 64, 78]. Work on traditional and hybrid craft similarly shows that expertise is transmitted through situated demonstrations, material engagement, and finely tuned bodily cues rather than purely verbal instructions [27, 84]. Building on these perspectives, we situate ViSTAR as a system that highlights embodied aspects of basketball skill learning by linking external feedback to how movements feel, while still operating on observable kinematic data. Concretely, ViSTAR translates kinematic deviations into feedback cues about posture, rhythm, and timing, anchored to drill phases and across-repetition comparisons, prompting athletes to self-check felt stability and weight transfer. Rather than directly measuring or modeling internal sensations, ViSTAR uses these kinematics-derived cues to prompt athletes' reflection on balance.

## 2.3 AI Sports Coaching System

AI is transforming the sports industry by enhancing performance analytics and decision-making [71, 80, 88]. This work focuses on the role of AI in sports coaching, addressing the limitations of traditional coaching resources, which are often restricted by cost, location, or availability. To fill this gap, researchers have explored AI-driven systems that provide automated or semi-automated feedback during training. Early systems were typically rule-based, relying on expert-encoded domain knowledge. For instance, Yin et al. [86] proposed a knowledge-based framework for intelligent team training. While interpretable, such systems lacked adaptability and scalability. Recent advances have shifted toward data-driven approaches that use sensors and machine learning to analyze performance and deliver feedback. ARRow [32], for example, is an AR system that provides real-time rowing feedback by visualizing biomechanical metrics such as stroke timing and posture using 3D skeletons. Similarly, PoseCoach [55] offers customizable video-based coaching by comparing novice and expert poses, presenting differences through 3D animations. These systems generally use computer vision to compare user motion to ideal models and visualize discrepancies for correction. More recently, Large Language Models (LLMs) have enabled adaptive, conversational coaching. GPTCoach [40], for

instance, combines wearable sensor data with motivational interviewing to deliver personalized activity plans. Building on these developments, our system integrates LLM-powered feedback with immersive visualizations. Unlike prior systems that rely solely on either visual or textual cues, our system offers a tightly integrated approach that combines immersive 3D motion visualizations with AI-generated verbal feedback. The verbal feedback offers timely, specific, and actionable guidance, helping users recognize posture errors, understand corrections, and refine their movements more effectively during practice.

### 3 FORMATIVE STUDY

We conducted a formative study to understand the current practices and challenges players encounter when learning and practicing basketball skills.

#### 3.1 Study Setup

**Participants:** We interviewed five university basketball players (mean age: 20.8 years, mean playing experience: 13.2 years) and two basketball coaches (mean age: 25.5 years, mean coaching experience: 5.5 years) to gain insights into their experiences and perspectives.

**Procedure:** Each session consisted of a 45-minute initial interview, a 15-minute skill training demonstration, and a 10-minute final interview. With consent to record, we asked participants about their background, training routines, self-learning methods, and challenges. After the interview, we showed three prepared skill training videos and asked participants to choose one to learn. They then practiced the skill while reviewing their captured motions. Finally, we conducted a post-training interview to understand their experience and how their perspectives evolved.

**Data Analysis:** We transcribed the audio recordings and applied Grounded Theory analysis [61, 62] to identify key difficulties. Two authors independently coded the transcripts and jointly reviewed the codes to establish categories. In cases of disagreement, the coders resolved differences through discussion. Inter-coder agreement, measured using Cohen’s Kappa, was 0.90.

#### 3.2 Findings Summary

The interviews revealed that participants’ skill-training process typically consists of three major components: 1) *Learning*: repeatedly watching video clips; 2) *Practices*: attempting to imitate the demonstrated movements; 3) *Improving*: recording and reviewing their own motions using tools such as loop playback and slow motion, then making adjustments based on perceived errors and how the motion felt. However, participants often struggled to translate a vague sense that something was “off”—for example, in their balance, timing, rhythm, or weight transfer—into concrete, body-part-specific corrections. We also identified three major difficulties in participants’ skill-learning process:

##### D1. Difficulty Understanding Skills from External Resources.

Video clips were the primary resource for learning, but participants noted that such materials are limited in perspective and lack multi-angle views. As one participant observed, “*Being able to pause and look at the play from multiple angles would be extremely helpful.*” Without varied viewpoints or 3D representations, participants struggled to perceive spatial structures, body alignment, timing,

and coordination [60]. Participants also reported that passive observation alone was insufficient, particularly for complex motor skills. As another remarked, “*It’s easier to understand when someone demonstrates in front of me.*” These findings highlight the need for guided instruction that breaks motions into smaller, digestible segments and emphasizes subtle but critical details such as joint articulation and timing accuracy.

**D2. Difficulty Recognizing Errors During Practice.** Many participants expressed a desire to compare their performance directly, side-by-side, with an expert model. One participant explained, “*Comparing my shot side-by-side with another could be insightful, as it would help me see the coordination of movements more clearly.*” Current tools, however, do not support synchronized comparisons. As a result, participants find it easier to detect mistakes when receiving external feedback. As one admitted, “*It’s hard to judge on my own. Having a more objective way to spot mistakes helps me understand better*” [22]. Without such references, their internal sense of whether a movement was “right” or “wrong” often remained vague. While some participants clip key moments to review with coaches, this support is not always available, leaving athletes to struggle with inefficient self-assessment.

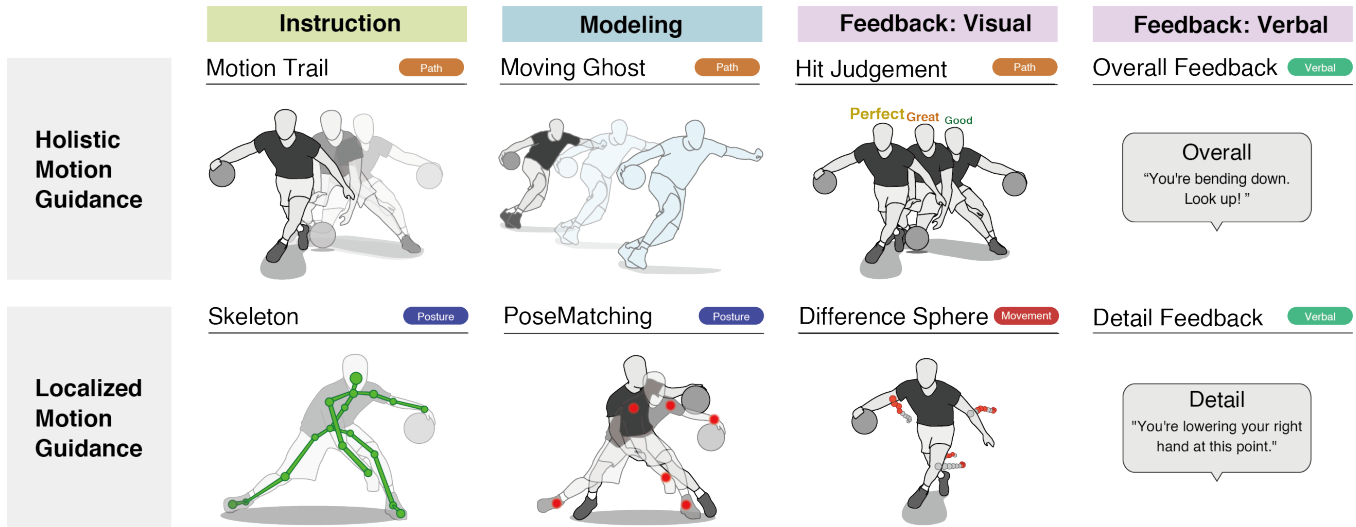
**D3. Difficulty Turning Intuition into Actionable Corrections.** Even when participants sensed that a movement felt wrong, they lacked the guidance to correct it. Participants often described how the move felt overall (e.g., off-balance, out of rhythm, or mistimed), but still struggled to pinpoint which body part to adjust, in what direction, and at what moment, particularly without expert input. As one participant explained, “*I couldn’t determine which parts I needed to improve, leaving me feeling uncertain.*” Prior work has shown that high-quality feedback boosts motivation, self-confidence, and training satisfaction [11]. Yet feedback quality varies with a coach’s expertise and communication style, and coaches are not always present. Participants noted that vague comments such as “*Be more aggressive*” left them confused, whereas precise instructions like “*Shift your weight to the right foot*” were more helpful. Overall, these findings point to the need for systems that can deliver clear, consistent, and context-specific feedback that connects how a move feels (e.g., balance, rhythm, timing, weight transfer) with concrete, interpretable cues about how to adjust one’s posture or coordination, independent of time, location, or coach availability.

## 4 ViSTAR: AR Coaching Agent for Behavioral Skills Training

To address the difficulties identified in our formative study, we designed ViSTAR, a AR system based on BST framework to provide both visual and verbal guidances to help users in basketball skill learning.

### 4.1 Usage Scenario

To illustrate the workflow of ViSTAR, we describe a hypothetical learner, Kooto, practicing a crossover dribble. First, ViSTAR presents the target motion as a 3D AR avatar, segmented into phases and viewable from multiple angles. Unlike conventional videos that restrict learners to a fixed viewpoint, the 3D AR avatar in ViSTAR can be viewed from multiple angles. Kooto can pause, rotate,



**Figure 2: Motion guidance strategies across BST stages. The top row represents Holistic Motion Guidance (e.g., Motion Trail, Moving Ghost, and Hit Judgement) that supports overall flow and coordination through path guidance. The bottom row shows Localized Motion Guidance (e.g., Skeleton, PoseMatching, Difference Sphere) focusing on joint-level feedback for fine-grained correction.**

and focus on subtle details such as joint articulation and timing. This flexibility enables him to examine the motion from different perspectives and gain a clearer understanding of complex movements, thereby addressing D1 by offering richer visual access to the motion. As he follows along, Kooto’s 3D pose is captured in real time and automatically aligned with the reference model. These visual overlays and synchronized side-by-side comparisons offer a consistent reference for self-assessment, helping Kooto observe subtle differences that can be overlooked, helping to mitigate D2 by making discrepancies more visually salient. Kooto then rehearses the skill by recording his own attempts, replaying them with expert overlays, and iteratively refining his motion. Finally, ViSTAR provides comprehensive multi-faceted feedback: visual heat maps emphasize regions where adjustments are beneficial, and verbal coaching offers precise and actionable suggestions such as “shift your weight to the right foot before crossing over.” This specificity aims to move beyond vague remarks (e.g., “be more aggressive”) by offering more targeted guidance that players can adapt to their own body and style, thus aiming to address D3 by offering more specific, coach-like cues. Through this workflow, ViSTAR operationalizes the Behavioral Skills Training (BST) framework of instruction, modeling, rehearsal, and feedback in an immersive AR environment.

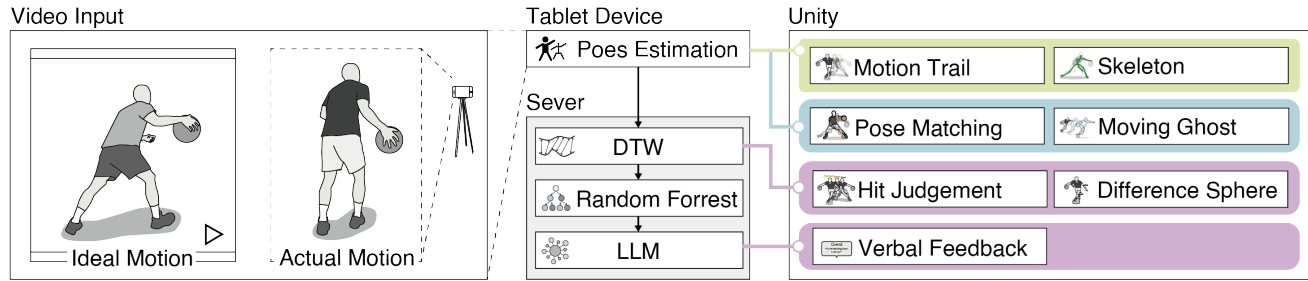
## 4.2 BST-grounded Guidance Design

The design of ViSTAR is grounded in the Behavioral Skills Training (BST) framework (Figure 2), a well-established method for improving skill performance across sports domains [29, 66, 70, 82]. BST consists of four stages: 1) *Instruction* – the trainer clearly describes the target skill, 2) *Modeling* – the trainer demonstrates the skill in detail, 3) *Rehearsal* – the learner practices the skill, and 4) *Feedback* – the trainer reviews the learner’s performance and suggests improvements. Beyond structuring error recognition, this framework also highlights embodied aspects of skill learning by organizing how learners move between attending to the overall feel of a movement

(e.g., rhythm and timing) and making specific postural adjustments, which informed our guidance design.

In our design, ViSTAR acts as **the trainer** by providing guidance in the *Instruction*, *Modeling*, and *Feedback* stages. To achieve this, we incorporated motion-guidance strategies identified in prior work [23], spanning two complementary dimensions: **Holistic guidance**, which emphasizes overall flow, timing, and full-body coordination (and can help learners notice changes in their balance across the move), directing users along intended trajectories; **Localized guidance**, which emphasizes accuracy in specific joints or body segments, such as highlighting key-frame poses to reinforce posture while accounting for individual differences (e.g., height, limb-to-torso ratio). We instantiated these strategies into guidance features for each stage as follows:

- **Instruction Guidance:** Skills are introduced through motion trails that visualize trajectories, combined with a 3D avatar and overlaid skeleton that highlights joint articulation (Figure 2a). This provides a clear reference of spatial and temporal structures, addressing D1.
- **Modeling Guidance:** Expert motions are demonstrated using moving ghost and a pose-matching interface (Figure 2b). The interface segments skills into discrete poses, providing further support for D1. Moreover, when the user follows the discrete poses, the system can provide real-time feedback. Specifically, when misalignment occurs, red spheres appear on the relevant joints and disappear once corrected, offering progressive, joint-level feedback on angles and timing. This partially addresses the error recognition challenges (D2).
- **Feedback Guidance (Visual):** In the feedback stage, ViSTAR combines holistic and localized guidance (Figure 2c) to address D3. For holistic guidance, a hit-judgment mechanism provides rhythm-game-style cues – “Perfect,” “Excellent,” “Great,” “Good,” or “Imprecise” – to reflect temporal



**Figure 3: Overview of the system workflow. User motion is analyzed using pose estimation, DTW, and Random Forest, and the resulting analysis is used to generate motion guidances, which are visualized in Unity through multi-faceted feedback.**

accuracy. This helps users refine timing and coordination according to their physical characteristics and preferred tempo. For localized guidance, ViSTAR synchronizes the learner’s motion with an expert reference in a side-by-side AR view. Misaligned joints are marked with red spheres that fade as alignment improves, providing immediate, interpretable feedback on spatial deviations. Together, these two visual feedback modes help learners understand both *when* and *where* their execution diverges from the reference.

- **Feedback Guidance (Verbal):** To further address the lack of actionable coaching, ViSTAR converts joint-level analysis into natural language suggestions (Figure 2d). These context-specific prompts mirror the effectiveness of precise coaching comments observed in our formative study (e.g., “Shift your weight to the right foot”), enabling iterative refinement during solo practice without requiring a coach to be present.

## 5 Technical Implementations

ViSTAR leverages AR’s spatial visualization and embodied interaction capabilities, together with the language generation capabilities of LLMs, to support the guidance introduced in Sec. 4. Here, we detail the key technical details and innovations for each stage. ViSTAR focuses on observable kinematics (e.g., posture, timing, and rhythm) and turns them into spatial and verbal cues that learners can align with their own bodily sensations. Our goal is not to fully model embodied experience, but to surface when balance, weight transfer, or tempo likely feel “off” and translate these patterns into concrete, coach-like guidance. To keep the hardware resource low and make deployment feasible, ViSTAR relies on a single front-facing RGB camera for 3D reconstruction. This design avoids the cost and calibration overhead of multi-camera setups. However, a multi-camera configuration could further improve joint reconstruction quality and robustness.

### 5.1 Instruction – 3D Reconstruction and Animate Avatar

To support the instruction phase and address D1, we integrate a monocular pose estimation model [3] that reconstructs full-body motion from a single 2D video stream. The model predicts 3D joint locations and rotation parameters from RGB frames, which we then retarget to a rigged avatar. This process transforms ordinary camera

input (e.g., from a smartphone) into a 3D representation of expert demonstrations as well as the user’s own performance. Users can select virtual coaches (e.g., their favorite players), which serve as consistent demonstrators. Using joint rotation data extracted with the model [3], we animate a 3D avatar that is viewable from any angle, reducing spatial ambiguity. To support understanding, the system overlays *motion trails* (path guidance) and *skeleton visualization* (posture guidance), as illustrated in Figure 2. Playback controls (pause, rewind, slow) let users inspect tempo and detail, revealing subtle joint and timing cues.

### 5.2 Modeling – Motion Step-by-Step Breakdown

To support skill understanding (D1), we implement a dynamic “moving ghost” visualization in the AR scene. The ghost is realized by animating a semi-transparent avatar with expert joint rotations for each frame. This path guidance goes beyond static frames, presenting the entire movement fluidly in time and space as a holistic guidance. For pose matching, ViSTAR normalizes joint positions by user-reported height so that expert and learner poses are represented in a common reference scale. Pose matching operates over segmented intervals (4 or 8 segments per skill) using both joint angles and positions. For each joint, we compute the angular deviation between the expert’s and learner’s rotation matrices, and the positional error as the Euclidean distance between normalized joint positions. A joint is considered aligned when the angular error is below  $30^\circ$  and the positional error is below 0.1 m, providing a coarse tolerance band that leaves room for individual rhythm while still flagging clearly problematic deviations. When at least 75% of joints in a segment meet these criteria, the system advances to the next segment. We do not perform a temporal alignment over the entire sequence, since the expert avatar already leads the learner through a scripted sequence of 4 or 8 segments. Real-time feedback is provided by overlaying red spheres on joints that exceed these thresholds. Markers fade as alignment improves, providing continuous correction cues. Crucially, these criteria do not enforce perfect replication, but promote adaptation: players internalize essential movement patterns and refine them to suit their own body characteristics and physical capabilities (e.g., anthropometrics, mobility, strength), focusing on effective, sustainable execution rather than identical form.

### 5.3 Feedback – Generating Multi-faceted Feedback

Unlike the segmented pose matching (subsection 5.2), where the learner rehearses one phase at a time under system-guided timing, the multi-faceted feedback operates on the full skill performed at the learner’s own pace. These recordings typically contain local tempo variations within the skill as well as extra frames before and after the core movement, so we use a DTW-based temporal alignment to fairly compare expert and learner trajectories and trim onset/offset noise over the entire sequence. We describe how ViSTAR implements multi-faceted feedback through hit judgement, spatial visual feedback, and verbal feedback. Once the video is captured, ViSTAR reconstructs the player’s 3D movement and uploads it to the server, where Dynamic Time Warping (DTW) [69] and motion alignment algorithms compute temporal and spatial correspondences with the expert reference (Figure 3). This alignment ensures that feedback remains consistent even when the learner’s execution differs in tempo or scale.

**5.3.1 Motion Alignment via Sliding Window Matching.** Since users manually start/stop recording, sequences include extraneous frames (e.g., button taps). We remove these irrelevant segments with a sliding window over the full sequence: for each window, we compute the DTW distance to the full expert motion and select the window with the minimum distance as the performed segment. This jointly optimizes start and length, making the method robust to speed variation:

$$t^* = \arg \min_{t \in [0, L_u - L_i]} \text{DTW}(M_{\text{user}}[t : t + L_r], M_{\text{ref}})$$

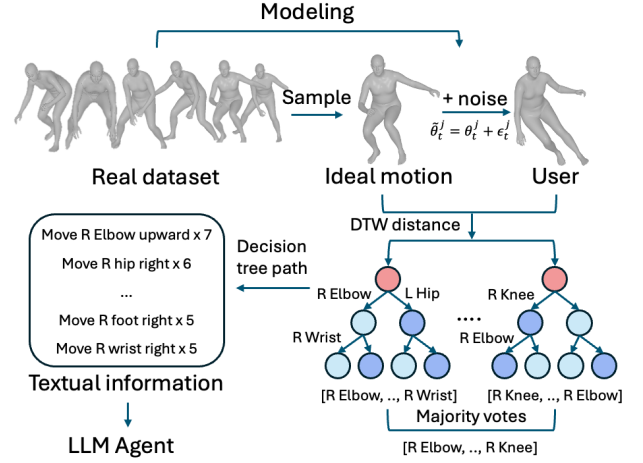
where  $L_i$  is the length of the ideal motion and  $L_u$  is the length of the user’s motion.

Furthermore, due to variations in player speed, the duration of user motions can be longer or shorter than the ideal. Using the best-aligned start index  $t^*$ , we perform variable-length matching by adjusting the window length  $L_w$  in the range  $[0.5L_i, L_u - t^*]$  and compute:

$$L_w^* = \arg \min_{L_w} \text{DTW}(M_{\text{user}}[t^* : t^* + L_w], M_{\text{ref}})$$

This approach ensures robustness to speed variations by selecting the best-aligned user segment based on both starting point and optimal length. All subsequent multi-faceted feedback (for hit judgement, difference sphere, and verbal feedback) are applied on the DTW-aligned trajectories.

**5.3.2 Hit Judgement and Spatial Visual feedback.** Temporal accuracy, timing and rhythm, is another critical yet hard to self-assess (D2). As a holistic guidance, our system implements a hit judgement mechanism over segmented intervals. The expert motion is divided into equal parts (e.g., 4 or 8 segments), and each segment is compared to the corresponding user motion by measuring its duration. Similar uniform time-normalization is widely used in biomechanics [20, 89], where movement cycles are rescaled to a common duration before comparing joint kinematics or other performance measures across trials. Timing accuracy is computed



**Figure 4: Verbal feedback generation pipeline. User motion is compared to ideal motion using DTW and a decision tree, and key misaligned joints are identified. An LLM generates feedback based on motion descriptors from the dataset.**

using the following formula:

$$\text{Score} = \max \left( 0, 100 - \left| \frac{T_{\text{ideal}} - T_{\text{actual}}}{T_{\text{ideal}}} \right| \times 100 \right)$$

Scores map to visual labels—*Perfect* ( $\geq 90\%$ ), *Excellent* ( $\geq 80\%$ ), *Great* ( $\geq 70\%$ ), *Good* ( $\geq 60\%$ ), *Imprecise* ( $\geq 50\%$ ), providing immediate cues on rhythm and speed for coach-free self-correction.

For spatial feedback, we render user and expert side-by-side in AR with synchronized playback. Red spheres highlight joints whose DTW-aligned positional error exceeds a distance threshold (we use the same 0.1, m tolerance as in the segmented phase), while gray ones denote joints within tolerance. *when* and *where* deviations occur. Combined with segment-level timing scores, this dual-layer feedback offers interpretable spatiotemporal cues that enable self-directed review and correction.

**5.3.3 Generating Verbal Feedback.** To help users identify and correct mistakes (D3), ViSTAR provides actionable verbal feedback. We convert both ideal and user numerical motion data into textual inputs for LLMs, which are not designed to handle raw joint angles. Without contextualization, LLMs struggle to interpret such data [50]. While some approaches leverage time-series structures [38], they require large-scale training. Instead, ViSTAR proposes a lightweight method combining dynamic time warping and random forest to identify and summarize key joint differences into concise textual feedback.

**Identifying different joints.** Euclidean distance is ill-suited for spatiotemporal motion. We compare Skinned Multi-Person Linear model (SMPL) [58]-based joint-rotation sequences (frames  $\times 24 \times 3$ ) using FastDTW [69]. When a joint-angle pair exceeds a threshold, it is verbalized (e.g., “move your left knee rightward; reduce excessive motion”), aggregated, and passed to an LLM to produce natural-language feedback. This naive listing can still surface too many joints, risking overload and diluting focus.

**Summarizing Important Motions using Random Forest.** To reduce excessive feedback, we use a Random Forest classifier

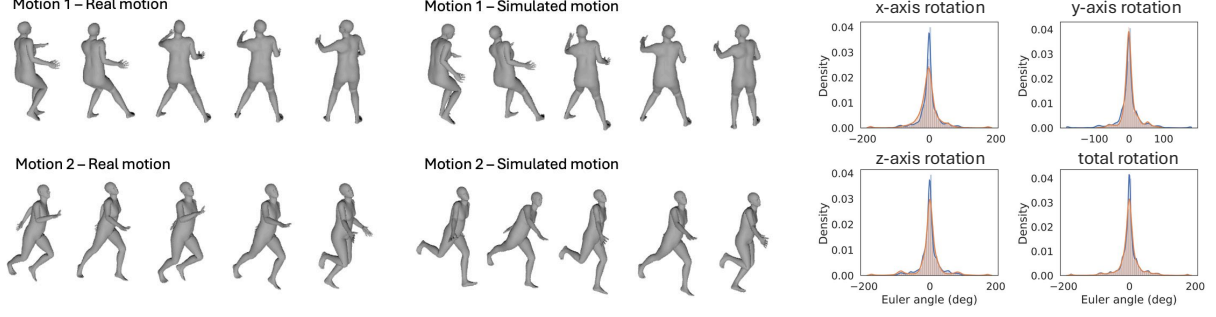


Figure 5: Comparison between real and simulated motions, along with distribution visualizations.

[9] to prioritize the most critical joint movements, focusing on informative and actionable cues rather than all joints. Random Forest, an ensemble method of decision trees, offers interpretable decision paths. By analyzing these paths (not model-level feature importance), we can identify which joint-angle differences most influenced the prediction, enabling instance-level interpretability. We selected Random Forest over deep neural networks or XGBoost [16] due to its transparency, lightweight nature, and robustness with limited data. Although XGBoost supports instance (e.g., level interpretation via SHAP values), SHAP [59] is computationally expensive and less intuitive to trace. Random Forest offers easier integration with LLMs and requires minimal tuning. We de-emphasize spine and torso joints in verbal feedback, since trunk posture and balance are easily visible in the full-body AR avatar and joint markers, whereas multi-limb coordination errors are harder to spot without explicit guidance. In line with prior systems that focus metrics on a set of limb joints [19, 23], our LLM prompts therefore primarily summarize recommendations at the multi-limb (arm/leg) level, while still surfacing spine-related cues when they are primary sources of error.

**Training with a Synthesized Dataset.** Since no dataset pairs expert and user motions for the same action, we synthesize one from public mocap (e.g., CMU MoCap [77]). We extracted 61 basketball-related motions (e.g., dribbling, shooting) and segmented them into 178 shorter sequences for finer variation then split 80%/20% for train/test. To simulate realistic errors, we added joint-level noise by computing frame-wise joint angle differences across time  $t$ , joint  $j$ , and axis  $a \in \{x, y, z\}$ , then sampling from noise distributions to create plausible user motion variants (see Figure 5).

In principle, the same synthetic perturbation strategy could be adapted to other motion datasets, but in this work we only apply and validate it on basketball-related skills. While generative models such as Motion Diffuse [87] can create motion variations, they have limited capacity to model the inter-personal variations for the same action. Instead, we model temporal differences in joint angles by fitting an exponential distribution to their absolute values. For each joint  $j$  and rotation axis  $a$ , we compute the temporal angle differences as:

$$\Delta\theta_t^{j,a} = \theta_t^{j,a} - \theta_{t-1}^{j,a}, \quad t = 2, \dots, T,$$

and fit an exponential distribution  $\mathcal{E}(\lambda_{j,a})$  to the set

$$\left\{ |\Delta\theta_t^{j,a}| \right\}_{t=2}^T \sim \mathcal{E}(\lambda_{j,a}).$$

This formulation captures how frequently and abruptly each joint tends to move over time. The parameter  $\lambda_{j,a}$  is estimated from the entire sequence, reflecting the typical magnitude of motion for the given joint and axis. Using the fitted distribution, we generate synthetic noise for each joint  $j$ , axis  $a$ , and time  $t$ :  $\epsilon_t^{j,a} =$

$$\epsilon_{\text{exp}}^{j,a} + \epsilon_{\text{norm}}^{j,a} \quad \text{where} \quad \begin{cases} \epsilon_{\text{exp}}^{j,a} \sim \text{Exponential}(\lambda_{j,a}) \\ \epsilon_{\text{norm}}^{j,a} \sim \mathcal{N}(0, \sigma_{j,a}^2) \end{cases} \quad \text{The exponen-}$$

tial component  $\epsilon_{\text{exp}}^{j,a}$  captures the typical magnitude and abruptness of joint movement, as estimated from the data. To better simulate the natural variability and randomness observed in human motion, we add a small amount of Gaussian noise  $\epsilon_{\text{norm}}^{j,a}$ , which introduces fine-grained stochasticity and avoids overly deterministic trajectories. Then, we create the noisy (user-like) motion:  $\tilde{\theta}_t^{j,a} = \theta_t^{j,a} + \epsilon_t^{j,a}$ . We define the input feature vector  $\mathbf{x}_t \in \mathbb{R}^{N \times 3}$  as the difference between the noisy and original motion:  $\mathbf{x}_t = \tilde{\theta}_t - \theta_t$

where  $N$  is the number of joints (e.g., 24), and each joint has 3 rotational angles. The corresponding label vector  $\mathbf{y}_t \in \{0, 1\}^N$  indicates which joints have been perturbed, where the  $j$ -th component  $y_t^j \in \{0, 1\}$  is 1 if noise is added to joint  $j$  at time  $t$  and 0 otherwise. The synthetic dataset in Figure 4 provides training pairs  $(\mathbf{x}_t, \mathbf{y}_t)$  for the Random Forest, where  $\mathbf{x}_t$  contains motion features from both the original and noise-perturbed sequences as inputs, and  $\mathbf{y}_t$  labels the joints to which noise was applied. We extract instance-level explanations by tracing the decision paths of the trained forest and summarize them as inputs to the LLM for verbal feedback as shown in section 5.4. In decision tree path, each node encodes explicit thresholds (e.g., “joint angle > value”), aligning with our goal of generating interpretable, condition-driven feedback. In contrast, SHAP requires extra processing to convert scores into language, making it less suitable for fast, lightweight feedback generation.

## 5.4 Implementation

We implemented four sequential steps: *Instruction* (3D avatar), *Modeling* (pose matching), *Rehearsal* (practice/recording), and *Feedback* (visual, timing, verbal). The system is built in *Unity* and deployed on a *Magic Leap AR headset* for immersive in-situ visualization. On the client side, *Unity* handles real-time rendering of 3D avatars, motion overlays, and AR feedback. For backend services, we use *FastAPI* to manage data processing pipelines and communication between the AR client and external servers. Motion data is processed by an efficient pose estimation model [3] that runs at 30 FPS on *iPhone 16* for real-time capture, while higher-accuracy models

can be integrated if latency is not critical. The verbal feedback module leverages the *OpenAI GPT-4 API*, with pre-processed motion descriptors and Random Forest decision paths provided as inputs to generate concise coaching feedback.

#### Prompt for Generating Feedback Summarization

You are an expert movement coach that helps users correct their body posture based on feedback from a motion analysis system. Your task is to take a list of movement instructions and summarize them in a **clear, natural, and structured** way that is easy for users to understand, while prioritizing the most important corrections.

#### Instructions:

- (1) **Group movement instructions by body region:**
  - **Upper Body:** shoulder, elbow, wrist → refer to as **arm**
  - **Lower Body:** hip, knee, ankle, foot → refer to as **leg**
  - Ignore spine/torso feedback unless it's the only area mentioned
- (2) **Identify the most critical issue** in each region:
  - Prioritize joints with **multiple or strong directional issues**
  - Focus on **excessive movement** that impacts balance or coordination
- (3) **Summarize feedback using only 1–2 concise sentences total:**
  - Mention only **one or two joints** maximum
  - Use language that is **specific, actionable, and understandable**
- (4) **Use user-friendly phrasing and avoid technical language:**
  - Instead of “Reduce movement right”, say “Try to limit movement to the right”
  - Instead of “Maintain movement”, say “Keep the movement as it is”

**Output Format:** Write a short paragraph that:

- Clearly describes the most important movement issues
- Includes a **suggestion for improvement**
- Avoids overloading the user with too many joint names or directions

#### Example Input:

Move your left knee left and Reduce movement.  
Move your left elbow down and Reduce movement.

#### Example Output:

- You may lose balance when lowering your left leg. Try adjusting your right leg as well.
- Your leg is moving in opposite directions. Try referencing a natural dribbling motion.

## 5.5 Computational Evaluation and Time Costs

**Computational Evaluation.** To assess the effectiveness of our verbal feedback module, we conducted a computational evaluation of the Random Forest classifier. We ran an ablation over two hyperparameters, *n\_estimators* and *max\_depth*, using our synthesized dataset (Sec. 5.3.3). Table 1 reports accuracy, precision, recall, and

F1-score across settings. The best configuration (*n\_estimators*=5, *max\_depth*=None) achieved average precision 0.8328, recall 0.9496, and F1-score 0.8874, which we adopt in our system. This performance demonstrates high recall to capture as many critical joint differences as possible while maintaining strong precision to avoid overwhelming users with excessive corrections.

To assess how realistic these synthetic pairs are relative to real expert motion, we compare the Euler-angle distributions of real and simulated poses, since no public dataset provides paired expert and learner executions for our target skills. Figure 5 visualizes the per-axis angle distributions for real versus simulated data, and Table 2b reports their Kullback–Leibler (KL) [46] and Jensen–Shannon (JS) [51] divergences. Both metrics measure how far two probability distributions deviate from each other, with lower values indicating a closer match. In our case, the synthetic poses remain statistically close to the real motion distribution, yet the injected perturbations still introduce differences from the original motions.

**Time Costs.** We implemented four sequential steps: Instruction (3D avatar), Modeling (pose matching), Rehearsal (practice/recording), and Feedback (visual, timing, verbal). Processing times on average are 0.31 s (pose estimation + visualization), 0.37 s (DTW), 0.01 s (random forest), and 2.3 s (LLM API). To support real-time feedback, we use an efficient pose estimation model [3] that runs at 30 FPS on iPhone 16, though higher-accuracy models can be integrated if needed.

## 6 USER STUDIES

We conducted two user studies with basketball players to evaluate verbal feedback accuracy and the usability and engagement of ViSTAR. The first study compared AI-generated verbal feedback with a real coach’s feedback after participants viewed both ideal and user videos. The goal of this study is to assess how participants perceive the quality of our coaching LLM’s verbal feedback compared to a human coach. The second study compared traditional practice (e.g., self-observation with video) and our AR system with a virtual coach, evaluating how each condition helped users identify and correct their errors. Both studies also assessed usability, engagement, and overall user experience.

### 6.1 Participants & Experiment Set-up

We recruited 16 basketball players (P1–P16; M = 16, Age: 20–34) via university mailing lists. Participants averaged 6.6 years of experience (range: 2–18) and were categorized as beginner (1), intermediate (12), and advanced (3). All had experience practicing alone or with peers. Since ViSTAR targets all levels, we intentionally recruited participants from a broad skill range to reflect diverse user needs. The study was conducted on an indoor court and lasted 60–75 minutes per participant. Each received a \$20 gift card. We selected five basketball skill videos from online coaching videos [7, 34, 68] and recorded corresponding performances to create five video pairs. These were used to obtain both expert and AI-generated feedback. Two actions (e.g., spin seal and cross over) were also used in the second study as the practice tasks.

**Table 1: Ablation Study on `n_estimators` and `max_depth` for Random Forest. “–” indicates that `max_depth` is set to None, meaning that the tree depth is unlimited.**

<code>n_estimators</code>	<code>max_depth</code>	Accuracy	Precision	Recall	F1-score
5	–	0.7500	<b>0.8328</b>	0.9496	<b>0.8874</b>
10	–	0.7500	0.8257	0.9460	0.8818
20	–	0.7778	0.8231	0.9622	0.8872
5	1	<b>0.8333</b>	0.7718	<b>0.9856</b>	0.8657
5	3	<b>0.8333</b>	0.8056	0.9766	0.8829
5	5	0.8056	0.8124	0.9658	0.8825

**Table 2: Average processing time per sequence (in seconds) and KL/JS divergence between real and simulated pose distributions.**

Component	Time (s)	Axis	KL(real    sim)	JS(real, sim)
Pose Estimation+Vis	0.31	X	0.1128	0.0260
DTW Alignment	0.37	Y	0.1709	0.0392
Random Forest Sum.	0.01	Z	0.0963	0.0219
LLM API Call	2.30	All (XYZ)	0.0663	0.0132

**(a) Average processing time per sequence (in seconds).****(b) Divergence between the empirical Euler-angle distributions of real and simulated poses.**

## 6.2 Study Procedure

**Introduction & Pre-survey (10 mins).** We provided the participants with an overview of the study and obtained the consent form, and collected background information (level of basketball skill and years of playing experience).

**Task 1: Verbal Feedback Evaluation (15 mins).** The first experiment compared AI-generated and real coach feedback. For AI-generated feedback, we included a baseline (without random forest) and our proposed method (with random forest). Each video pair was paired with three feedback versions: real coach, baseline AI, and proposed AI in randomized order. To reduce bias, participants first completed a verbal quality assessment based on three criteria from our formative study: *clarity*, *identifiability*, and *actionability*. Questionnaires are in Appendix A.1. To ensure fairness and eliminate ordering bias, three verbal feedback types, (1) real coach, (2) baseline (DTW + LLM), and (3) our proposed Random Forest + LLM, were shown in randomized order per video, with participants unaware of their source.

**Task 2: Skill Training (35 mins).** To mitigate novelty effects, we intentionally placed the verbal feedback evaluation (Task 1) before the skill training task. In the second experiment, participants practiced two basketball actions under two conditions: baseline (self-observation) and ViSTAR (AR-based virtual coaching). A within-subjects design with Latin square randomization was used to balance order and task difficulty. Both conditions were presented in the same AR headset. In the baseline, participants watched an ideal performance, practiced twice, and reviewed their recordings to self-correct. In the ViSTAR, they received step-by-step training and multi-modal feedback from a virtual coach.

**Post-survey (10 mins).** After completing the two studies, participants were asked to rate the usability, engagement, and overall user experience of ViSTAR based on key factors [17, 48, 54]. In

addition, they were asked to compare the two conditions in terms of how effectively they could identify and correct their mistakes. Participants also provided individual evaluations of each feature used in the system.

## 6.3 Results

**6.3.1 Verbal Feedback Preference Results.** As shown in Table 3, our method received the highest number of first-place rankings (33 out of 80, 41.2%) and the fewest third-place rankings (7 out of 80, 8.8%), indicating a consistent overall preference. In contrast, the real coach feedback and baseline were selected first only 28.8% and 30.0% of the time, respectively, and received substantially higher third-place rankings (40.0% and 51.2%). These results suggest that our model-generated feedback was perceived as more understandable, identifiable, and actionable—criteria directly derived from our formative study. RF feedback received the highest number of first-place rankings, but a closer look at individual skills shows that preferences varied by motion, with some actions favoring baseline or coach feedback instead (Table 4).

**Table 3: Number and percentage of 1st, 2nd, and 3rd place votes for each verbal feedback type.**

Feedback Type	1st Place	2nd Place	3rd Place
Real Coach	23 (28.8%)	16 (20.0%)	32 (40.0%)
Baseline	24 (30.0%)	24 (30.0%)	41 (51.2%)
RF (Ours)	<b>33 (41.2%)</b>	<b>40 (50.0%)</b>	<b>7 (8.8%)</b>

**6.3.2 Skill Training Performance Comparison: ViSTAR vs. Self-Observation.** We analyzed the accuracy of users’ reproduced poses under two conditions: a traditional self-observation baseline and

**Table 4: Percentage of 1st-place rankings for each feedback type per motion.**

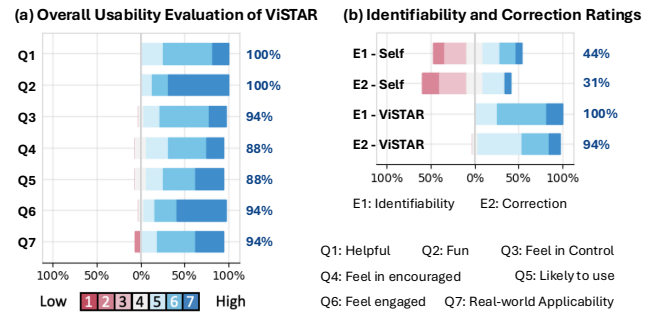
Motion	Real Coach	Baseline	RF (Ours)
CrossOver	18.8%	18.8%	<b>62.4%</b>
CrossSnatch	37.5%	18.8%	<b>43.7%</b>
ShammgodCissors	18.8%	<b>43.7%</b>	37.5%
SpinSeal	31.3%	<b>37.4%</b>	31.3%
TimHardaway In-And-Out	<b>37.4%</b>	31.4%	31.4%

our system, ViSTAR. Accuracy was measured as the average angular deviation (in degrees) between the user’s joint rotations and the reference motion. Each participant performed two trials per condition. In the first trial, participants using ViSTAR showed lower angular deviation compared to those using self-observation ( $M = 10.80^\circ$  vs.  $M = 14.95^\circ$ ). A Wilcoxon signed-rank test indicated a statistically significant difference between the two conditions ( $W = 22.0$ ,  $p = 0.0155$ ). In the second trial, participants again showed lower angular errors with ViSTAR ( $M = 11.32^\circ$ ) compared to the self-observation baseline ( $M = 14.60^\circ$ ). The Wilcoxon signed-rank test similarly revealed a statistically significant difference ( $W = 21.0$ ,  $p = 0.0131$ ).

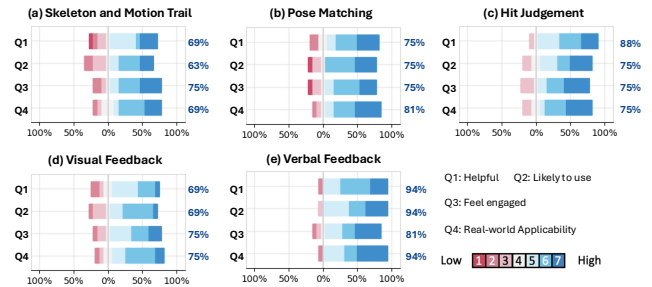
**ViSTAR felt more helpful for error recognition and correction.** As shown in the right panel of Figure 6, only 44% and 31% of participants could identify and correct errors using self-observation. In contrast, 100% and 94% of participants reported that ViSTAR helped them both recognize what went wrong and understand how to fix it. This improvement directly addresses D2 (error recognition) and D3 (receiving actionable feedback). The combination of modeling, visual feedback, and guided practice enabled users to better understand and internalize correct motion.

Participants highlighted that side-by-side avatar comparison in AR made it easier to detect mistakes, as in P3’s comment: “*easy to follow the movements because the AR scene closely resembled a real basketball environment.*” Visualizations such as red spheres and 0.25x slow motion helped users inspect details and assess themselves more effectively. For D3 (receiving actionable feedback), verbal feedback increased confidence. P4 noted “*I can trust and correct my motion when my own judgment and verbal feedback are the same.*” Beyond correction, it also prompted reflection. P10 remarked, “*The coaching saying is a new way to think about another aspect while training.*” While one user found some text “too specific,” most described the system as “engaging,” “helpful,” and “great at helping me analyze my movement.” In addition, a few participants reported confusion when their own judgment conflicted with the model’s feedback. In such cases, they sometimes hesitated or felt uncertain about which to trust. Nonetheless, no participants mentioned reconstruction errors as a source of distraction, suggesting that the current accuracy of our pose estimation was sufficient for training purposes. Notably, no participants reported failures of the pose estimation model or raised concerns about system malfunction due to reconstruction errors.

**6.3.3 Tool Evaluation of ViSTAR.** Figure 6 and Figure 7 present the results from the second task of our user study, where participants



**Figure 6: Overall usability ratings of ViSTAR across seven dimensions (e.g., helpfulness, engagement, applicability), showing high user satisfaction. (b) Comparison of identifiability and correction ratings between the self-observation baseline and ViSTAR, with ViSTAR receiving consistently higher scores.**



**Figure 7: Feature-level evaluation of five feedback components: (a) Skeleton and Motion Trail, (b) Pose Matching, (c) Hit Judgement, (d) Visual Feedback, and (e) Verbal Feedback. User responses reflect consistently high scores across helpfulness, willingness to use, engagement, and applicability.**

experienced both self-observation (baseline) and ViSTAR conditions while practicing individual basketball motions. We collected Likert-scale ratings on 1) usability, engagement, 2) identifiability, correction support, and 3) feedback on individual system components.

**ViSTAR shows high usability and real-world applicability.** As shown on the left side of Figure 6, ViSTAR received high ratings across all usability criteria: helpful (Q1), fun (Q2), feel in control (Q3), encouraged (Q4), likely to use (Q5), feel engaged (Q6), and real-world applicability (Q7). Both Q1 and Q2 received 100% positive ratings (scores 5–7), while the remaining items also showed positive ratings, ranging from 88% to 94%. These results indicate that ViSTAR offers an intuitive and engaging experience with promising practical value. Open-ended feedback echoed these findings. Participants appreciated the multi-angle views and 3D avatar visualization. For example, P9 said, “*I can see from every angle, this is extremely awesome!*”, while others noted improved perception of spatial relationships and coordination. These immersive visuals helped them perceive spatial relationships and coordination more clearly than when reviewing flat video.

**Each core feature contributes to D1–D3 with distinct strengths and some limitations.** Figure 7 breaks down user responses across the five features in terms of helpfulness, engagement, and real-world applicability. (a) Skeleton and Motion Trail and (b) Pose Matching supported D1 by clarifying complex motions. Skeleton and Motion Trail helped users “*see every posture step of movement*” but could be visually dense. P8 found it “*too cluttered*” (scores 63–75%), suggesting a keyframe or pause-step approach to reduce load. Pose Matching (75–81%) aided spatial understanding. P2 valued its “*directed feedback*”, though P10 found it “*hard to follow during dynamic motions*”. However, some users noted that these features required effort to interpret during dynamic play, indicating that their utility may be partly tied to the novelty of seeing detailed overlays rather than sustained usability.

To address D2, (c) Hit Judgement and (d) Visual Feedback provided indications of correctness. Hit Judgement helped users pinpoint when their movement aligned or fell out of sync with the expected timing, with P6 noting that it “*showed where your movement was good and where it was not*” in relation to the intended timing. Visual Feedback was also reported as helpful (69–75%), with participants describing it as “*the easiest to understand where the problem was*” and “*I could actually compare my movement to the ideal.*” Still, when multiple red spheres appeared, some users reported difficulty focusing their attention. Yet a few participants found the cues difficult to map to specific motion phases, indicating that while generally helpful, some aspects require refinement for clarity and pacing.

(e) Verbal Feedback played a particularly strong role in addressing D3 by offering clear and actionable guidance. It received high positive ratings across items. Users praised it as “*The feedback is reasonable and coherent and concise. It’s easy to understand.*”(P5). Still, one participant felt there was too much to read and suggested, “*I really liked the verbal feedback! Maybe reduce the text a bit and use bullet points,*” implying that a more concise summary format could enhance usability. Taken together, these modules offered complementary strengths across D1–D3. While users responded positively to all components, common themes around visual clutter and pacing suggest opportunities to enhance clarity and reduce overload.

## 7 DISCUSSION AND FUTURE WORK

We position ViSTAR primarily as an enabling AR+AI platform and discuss the design implications and insights about how such systems can support bodily skill learning and self-guided practice.

### 7.1 Multimodal Feedback as a Key Enabler for AR+AI Coaching

**Verbal Feedback as a Design Pattern for Motion-to-Language Interfaces.** Although our evaluation centered on basketball skills, the verbal feedback pipeline of ViSTAR highlights how low-level spatio-temporal joint features can be translated into accessible, actionable coaching through natural language. It illustrates one promising way XR systems can connect motion data and language interfaces. Participants described verbal feedback with 3D avatar as “*felt like there is a live coach,*” reflecting a perceived sense of guidance that echoes prior reports of increased self-awareness

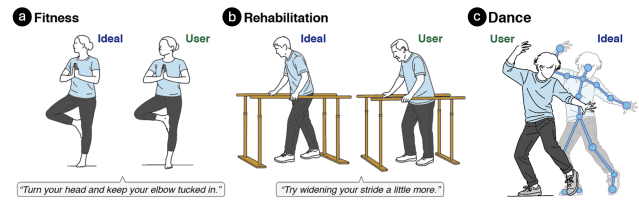
in AI-supported practice [81]. At the same time, concerns about authenticity and trust remained. One participant (P4) expressed skepticism, stating, “*I’m not sure I fully trust the feedback since it’s AI-generated,*” and felt it resembled a general chatbot. The speaking style also lacked the tone of actual athletes, creating dissonance. P4 noted that real coaches’ feedback tended to be high-level and strategic (e.g., “focus on timing”), whereas AI feedback was more detailed and specific (e.g., “shift your weight to the right foot”). Such trade-offs suggest a broader design principle: balancing precision and personalization in human–AI collaboration, where human coaches provide strategic guidance while AI systems contribute detailed, consistent feedback [47, 57, 65]. Together, these comments point to a specificity–authenticity tension in motion-to-language coaching: increasing kinematic specificity may improve actionability, yet can reduce perceived coach-likeness when tone and granularity diverge from coaching norms. These observations suggest that future systems may need to co-tune *what* is said and *how* it is said. (1) Highly specific, corrective language can be actionable, yet feel less “coach-like” when it mismatches users’ expectations of coaching tone and level. One practical direction is to provide strategic, coach-like cues by default, with optional drill-down details on demand. (2) To reduce “chatbot-like” impressions and support trust, verbal feedback should stay accountable to motion evidence—for instance, by briefly grounding claims in the avatar (e.g., highlighting the referenced body segment) and allowing lightweight verification such as replaying the relevant moment or a simple before/after comparison.

**Multi-modal Feedback: Beyond Multi-faceted Feedback.** To further support a sense of authenticity and user trust, future systems can go beyond multi-faceted feedback (e.g., verbal + visual) by integrating multi-modal input channels such as voice commands, gaze tracking, and contextual performance history. These modalities allow the system to infer the user’s intent, attention, and experience level, and in turn, tailor feedback dynamically. In current AR training system, users must interrupt their physical activity to operate AR controllers, for example, to view feedback, switch perspectives, or advance to the next segment. These interruptions may break immersion and impose physical and cognitive load. By integrating voice commands, users could interact with the system hands-free (e.g., saying “show me again” or “please recoding”), allowing for a more seamless and natural training experience. Such multimodal interaction could greatly enhance the usability and fluidity of AR-based coaching systems by reducing the friction we observed when participants had to interrupt movement to operate controllers. Evaluating how such modalities affect perceived usability and trust is an important direction for future work. Another opportunity lies in generating feedback using vision or video-language models that directly operate on images or videos of the motion. Exploring these end-to-end motion-to-language generation represents a promising direction for future work. This suggests multimodality is not merely a convenience feature, but a mechanism for managing intervention timing under ongoing physical activity. These observations suggest that multimodal input should protect practice continuity: (1) hands-free controls (e.g., voice) can reduce mid-drill interruptions for replays or view changes; and (2) signals like gaze or performance context may help time interventions (speak, defer, or wait for user pull) to avoid over-interrupting.

## 7.2 Design Challenges in Dynamic Motion: Balancing Visual Clarity and Feedback Precision

**Designing for Feedback in Dynamic Motion.** Prior work [85] demonstrated that 3D trajectory cues are effective for relatively linear movements (e.g., shooting, skiing). However, our findings reveal that in dynamic, high-frequency tasks like dribbling, these cues become cluttered and harder to interpret. Overall, participants found the trajectory trails helpful for understanding motion flow. However, several users reported difficulty interpreting overlapping visual elements, describing the trails as cluttered or vague during fast-paced sequences. This suggests that the effectiveness of trajectory-based feedback diminish in complex motions unless the visualizations are carefully designed to reduce cognitive load. Our hit judgement feature was valued for real-time feedback on timing and quality, but some users were confused by *“too many cues in rapid succession.”* These issues highlight a common tension in motion feedback: detailed feedback can be useful, but without clear localization or temporal anchoring, it risks overwhelming users. Future designs could benefit from adaptive visual abstraction strategies such as keyframe simplification, progressive disclosure, or context-sensitive highlighting to reduce overload while maintaining instructional clarity. This suggests two design directions: (1) reduce cue density through adaptive abstraction (e.g., keyframe simplification or progressive disclosure), and (2) improve interpretability by anchoring feedback to moments users can localize (e.g., event-/rhythm-based segmentation with context-sensitive highlighting). Although we currently divide expert motion into equal-length segments for hit judgment, consistent with common time-normalization practices in biomechanics analysis [20, 89], this simplification does not explicitly follow the natural rhythm or event structure of the skill. Future work should investigate event- or rhythm-based segmentation schemes that better align feedback with how athletes perceive phases of movement and balance changes.

**Balancing Feedback in Dynamic Motion Across Expertise Levels.** Our user study showed that the effectiveness of motion feedback is not uniform across expertise levels. Less experienced players often asked for simpler, more stepwise cues. For example, one 5-year player (P13) felt that the verbal feedback was *“too detailed”* and requested *“one-by-one suggestions,”* while another (P8) proposed that *“high-level comments or bullet points”* would make it easier to act on. In contrast, more experienced players with 10–18 years of basketball experience appreciated richer, more diagnostic guidance: one participant (P5, 18 years) noted that the verbal feedback provided *“detailed comments which are reasonable and coherent,”* and another (P10) emphasized that it helped them *“easily understand exactly what was going wrong.”* These qualitative patterns echo prior work showing that novices benefit more from explicit, directive feedback, whereas advanced learners benefit from more facilitative, reflective guidance and can better handle detailed information [72, 73]. From a design perspective, these observations raise questions about *scaffolded and responsive feedback strategies* that evolve with the learner’s trajectory. A natural direction for future work is to calibrate the granularity and framing of feedback as users progress, shifting from directive to more reflective guidance and examining how this affects learning outcomes. Such



**Figure 8: Beyond sports, our system could be applied to diverse domains by comparing user movements against ideal references and providing targeted feedback: fitness (a), rehabilitation (b), and dance (c).**

strategies could also be informed by user states such as cognitive load, confidence, or engagement, enabling systems to serve as interactive learning companions rather than static feedback channels, as suggested by prior work on action execution and observational learning [48, 90]. A practical implication is to scaffold feedback over time: (1) default to concise, stepwise cues for novices with optional drill-down, and (2) provide richer diagnostics for advanced users while monitoring overload (e.g., via interaction signals or self-reported difficulty).

## 7.3 Situating Motion Learning: Viewpoint, Social Context, and Beyond Basketball

**Viewpoint Alignment and Spatial Understanding in Motion Learning.** Traditional 2D video demonstrations are mirrored, requiring users to mentally reverse left and right when imitating movements. This can cause spatial confusion, particularly in lateral or asymmetric tasks. Prior studies [21, 35, 85] have suggested that such mirrored views hinder motor learning by increasing cognitive load. In designing our system, we deliberated between a mirrored (coach-facing) or egocentric (user-aligned) avatar view. We adopted the egocentric approach to minimize mental transformation, and participants reported no confusion with left-right alignment. This insight offers valuable implications for future instructional systems in sports, dance, rehabilitation, and other embodied domains. Understanding how different viewpoints (mirrored vs. aligned) shape perception and learning could inform the design of more effective training environments, especially for spatially complex skills. We encourage future studies to empirically compare these visualizations and quantify their impact on learning outcomes across different levels of user expertise.

**Design Implications for Social and Competitive Learning.** Our system focused on solo skill training and did not capture the reactive and interactive nature of real gameplay [17, 75, 76]. Participants trained in isolation without the presence of defenders or teammates, limiting opportunities to develop decision-making and adaptability under pressure. Real-world performance depends not only on technical execution but also on how players respond to others’ movements in dynamic, contested spaces. Future systems could move beyond isolated drills by introducing interactive avatars that simulate opponents or teammates, e.g., dribbling past a defender, coordinating with a teammate, or reacting to tactical shifts. Such designs would better support situational awareness, anticipation, and timing, which are central to competitive play. More broadly,

embedding reactive agents into training environments may foster collaboration and strategic thinking, bridging the gap between controlled practice and authentic gameplay. These opportunities extend beyond sports to domains such as dance, rehabilitation, and physical education, where spatial coordination and interpersonal responsiveness are equally critical.

**Generalization to Motion Learning in Other Domains.** Our current evaluations, datasets, and thresholds are tailored to basketball-specific, isolated motions. However, the same motion-to-language pipeline (i.e., extracting low-level joint data, mapping them to joint descriptors, and generating natural-language feedback) may generalize to other multi-limb coordination tasks in domains such as fitness, rehabilitation, or dance [33], as conceptually illustrated in Figure 8. Extending the approach would require re-calibrating biomechanical tolerances, redefining task goals in collaboration with domain experts [53], and validating the system with new participant populations.

## 7.4 Limitations

We discuss key limitations of our current implementation and study, many of which relate to ecological validity.

**Tracking Setup and Alternative Sensing Approaches.** Our study used a single player in a front-facing capture configuration with one RGB camera and an AR HMD. This setup aligns with our target use case of on-court individual skill training before games, where players typically perform short, isolated technical drills alone or with a coach rather than full, dynamic gameplay. Similar to prior AR/VR sports training work [14, 17, 24, 28, 37, 52], it enabled controlled evaluation of isolated skills. However, it does not capture multi-player or crowded scenarios, where player-player and ball occlusions would be more frequent and the HMD’s limited field of view would further constrain ecological realism. In addition, our current prototype lacks a dedicated ball-tracking module and does not model defenders or teammates, which prevents full alignment between ball trajectory and the 3D avatar, limits analysis of ball-body relations (e.g., release timing, catch alignment) and multi-player scenarios, and restricts claims about decision-making under pressure.

An efficient alternative to our external video-based tracking is to estimate body pose directly from AR headsets using on-device sensing (e.g., inside-out cameras and IMUs), which could reduce setup burden and improve portability for on-court drills. However, current headset-only approaches may provide limited full-body fidelity under fast sports motions and frequent occlusions. In the present work, we chose a single external RGB camera (e.g., a smartphone) to keep the setup lightweight and accessible using readily available, off-the-shelf equipment, without specialized sensors or additional instrumentation.

**Study Duration and Headset Comfort.** Each participant completed the training session within a single day, which constrained the amount of time they could spend practicing, adapting to the system, and incorporating the feedback. Although the system provides a highly immersive experience, the current HMD is relatively heavy. Therefore, we did not study longer, sweat-heavy, high-speed basketball sessions, where comfort, robustness, safety, occlusion,

field-of-view, and ball-interaction issues with head-mounted displays may become more pronounced.

With 16 participants, our results should be interpreted as providing exploratory trends rather than strong statistical claims. Future longitudinal studies with larger samples, repeated sessions on-court deployment in more realistic training scenarios, and lighter AR form factors (e.g., glasses-style devices) are needed to better capture how the system supports learning progression and sustained performance improvement. Our findings should be interpreted as exploratory evidence from a short, single-session study in a constrained setting, rather than as proof of training gains in real-world basketball practice.

**Simplified Feedback Features and Perturbation-Based Modeling.** In our current design, we de-emphasize spine and torso joints in verbal feedback, prioritizing limb coordination (e.g., foot placement, hand timing), which participants found harder to notice without explicit cues. Similar metric choices appear in prior systems that focus error descriptors on limb joints [19, 23]. However, this simplification also limits how fully the system can address whole-body coordination and balance, where torso angle plays a critical role. Incorporating richer torso descriptors and linking them to bodily sensations is an important direction for future work. Our Random Forest classifier is trained on synthetic motion pairs generated by injecting joint-level perturbations into expert motions. As quantified in Sec. 5.5, these simulated poses are close to the real expert motion distribution, but they still provide only a first approximation to how real learners move and make mistakes. Incorporating collecting paired expert-learner motion data that capture realistic error patterns are important directions for future work.

## 8 CONCLUSION

We presented ViSTAR, an AR training system that integrates 3D motion reconstruction with LLM-based verbal feedback to support self-guided athletic practice. ViSTAR uses the Behavioral Skills Training (BST) framework to organize visual and verbal feedback. In our study, ViSTAR yielded lower angular errors than traditional self-observation, but we treat these performance results as exploratory. Participants generally preferred AI-generated feedback over real coaches’, citing its clarity and actionability, and rated ViSTAR as usable and engaging. Verbal and visual feedback were especially helpful for noticing and interpreting errors. Overall, our results highlight the promise of multi-faceted feedback for motor skill learning and suggest design considerations such as egocentric viewpoints and adapting feedback to user expertise. We position ViSTAR as an enabling AR+AI platform that supports reflection on embodied aspects of skill learning rather than a fully validated training intervention.

## Acknowledgments

This work is supported by NSF grant CRCNS-2309041 and Harvard Data Science Initiative Trust in Science Fund Award.

## References

- [1] Anna Akbaş, Wojciech Marszałek, Anna Kamiernarz, Jacek Polechoński, Kajetan J Słomka, and Grzegorz Juras. 2019. Application of virtual reality in competitive athletes—a review. *Journal of human kinetics* 69 (2019), 5.

- [2] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 311–320.
- [3] Yukihiro Aoyagi, Shigeki Yamada, Shigeo Ueda, Chifumi Iseki, Toshiyuki Kondo, Keisuke Mori, Yoshiyuki Kobayashi, Tadanori Fukami, Minoru Hoshimaru, Masatsune Ishikawa, et al. 2022. Development of smartphone application for markerless three-dimensional motion capture based on deep learning model. *Sensors* 22, 14 (2022), 5282.
- [4] Richard Bailey and Angela Pickard. 2010. Body learning: examining the processes of skill learning in dance. *Sport, education and society* 15, 3 (2010), 367–382.
- [5] Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.* 59, 1 (2008), 617–645.
- [6] Jürgen Bernard, Nils Wilhelm, Björn Krüger, Thorsten May, Tobias Schreck, and Jörn Kohlhammer. 2013. Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2257–2266.
- [7] Zero Bounce. [n. d.]. TOP 3 BASKETBALL MOVES TO FOR REAL GAMES. <https://www.youtube.com/shorts/2zWWVvFjmew>. Accessed: 2025-04-09.
- [8] Nataliya Braun and Yasuhiro Kotera. 2022. Influence of dance on embodied self-awareness and well-being: An interpretative phenomenological exploration. *Journal of Creativity in Mental Health* 17, 4 (2022), 469–484.
- [9] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
- [10] Wolfgang Büschel, Anke Lehmann, and Raimund Dachselt. 2021. Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.
- [11] Joëlle Carpentier and Geneviève A Mageau. 2016. Predicting sport experience during training: The role of change-oriented feedback in athletes' motivation, self-confidence and needs satisfaction fluctuations. *Journal of sport and exercise psychology* 38, 1 (2016), 45–58.
- [12] Jacky Chan, Howard Leung, Kai Tai Tang, and Taku Komura. 2007. Immersive performance training tools using motion capture technology. In *Proceedings of the First International Conference on Immersive Telecommunications*. 1–6.
- [13] Jacky CP Chan, Howard Leung, Jeff KT Tang, and Taku Komura. 2010. A virtual reality dance training system using motion capture technology. *IEEE transactions on learning technologies* 4, 2 (2010), 187–195.
- [14] Hao Chen, Yujia Wang, and Wei Liang. 2022. Vcoach: Enabling personalized boxing training in virtual reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 546–547.
- [15] Po-Jung Chen, I-Wen Penn, Shun-Hwa Wei, Long-Ren Chuang, and Wen-Hsu Sung. 2020. Augmented reality-assisted training with selected Tai-Chi movements improves balance control and increases lower limb muscle strength in older adults: A prospective randomized trial. *Journal of Exercise Science & Fitness* 18, 3 (2020), 142–147.
- [16] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [17] 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*. 1–14.
- [18] Jean Côté, Joseph Baker, and Bruce Abernethy. 2007. Practice and play in the development of sport expertise. *Handbook of sport psychology* 3, 184-202 (2007).
- [19] Henrique Galvan Debarba, Marcelo Elias De Oliveira, Alexandre Lädermann, Sylvain Chagué, and Caecilia Charbonnier. 2018. Augmented reality visualization of joint movements for rehabilitation and sports medicine. In *2018 20th Symposium on Virtual and Augmented Reality (SVR)*. IEEE, 114–121.
- [20] Christopher A DiCesare, Adam W Kiefer, Scott Bonnette, and Gregory D Myer. 2020. High-risk lower-extremity biomechanics evaluated in simulated soccer-specific virtual environments. *Journal of Sport Rehabilitation* 29, 3 (2020), 294–300.
- [21] Florian Diller, Nico Henkel, Gerik Scheuermann, and Alexander Wiebel. 2025. SkillAR: omnipresent in-situ feedback for motor skill training using AR. *Virtual Reality* 29, 1 (2025), 1–11.
- [22] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106.
- [23] Hesham Elsayed, Kenneth Kartono, Dominik Schön, Martin Schmitz, Max Mühlhäuser, and Martin Weigel. 2022. Understanding Perspectives for Single-and Multi-Limb Movement Guidance in Virtual 3D Environments. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. 1–10.
- [24] Kazuhiro Esaki and Katashi Nagao. 2024. An efficient immersive self-training system for hip-hop dance performance with automatic evaluation features. *Applied Sciences* 14, 14 (2024), 5981.
- [25] Charles Faure, Annabelle Limballe, Benoit Bideau, and Richard Kulpa. 2020. Virtual reality to assess and train team ball sports performance: A scoping review. *Journal of sports Sciences* 38, 2 (2020), 192–205.
- [26] Hugh HK Fullagar, Liam D Harper, Andrew Govus, Robert McCunn, Joey Eisenmann, and Alan McCall. 2019. Practitioner perceptions of evidence-based practice in elite sport in the United States of America. *The Journal of Strength & Conditioning Research* 33, 11 (2019), 2897–2904.
- [27] Connie Golsteijn, Elise Van Den Hoven, David Frohlich, and Abigail Sellen. 2014. Hybrid crafting: towards an integrated practice of crafting with physical and digital components. *Personal and ubiquitous computing* 18, 3 (2014), 593–611.
- [28] Ut Gong, Hanze Jia, Yujie Wang, Tan Tang, Xiao Xie, and Yingcai Wu. 2024. VolleyNaut: Pioneering immersive training for inclusive sitting volleyball skill development. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 1022–1032.
- [29] Mairissa Harris, Laura Baylot Casey, James N Meindl, Douglas Powell, William C Hunter, and Diana Delgado. 2020. Using behavioral skills training with video feedback to prevent risk of injury in youth female soccer athletes. *Behavior Analysis in Practice* 13, 4 (2020), 811–819.
- [30] Nicola J Hodges and Ian M Franks. 2002. Modelling coaching practice: the role of instruction and demonstration. *Journal of sports sciences* 20, 10 (2002), 793–811.
- [31] Sebastian Hubenschmid, Jonathan Wieland, Daniel Immanuel Fink, Andrea Batch, Johannes Zagermann, Niklas Elmqvist, and Harald Reiterer. 2022. Relive: Bridging in-situ and ex-situ visual analytics for analyzing mixed reality user studies. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [32] Elena Iannucci, Chen Zhu-Tian, Iro Armeni, Marc Pollefeys, Hanspeter Pfister, and Johanna Beyer. 2023. ARrow: A real-time AR rowing coach. *EuroVis 2023-Short Papers* (2023), 73–77.
- [33] 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.
- [34] ILoveBasketballTV. [n. d.]. Top 5 Moves All Basketball Players Should Know GET SHIFTY! <https://www.youtube.com/watch?v=CGRJdBA6hGc>. Accessed: 2025-04-09.
- [35] Yasuyuki Inoue and Michiteru Kitazaki. 2021. Virtual mirror and beyond: The psychological basis for avatar embodiment via a mirror. *Journal of Robotics and Mechatronics* 33, 5 (2021), 1004–1012.
- [36] Sujin Jang, Niklas Elmqvist, and Karthik Ramani. 2015. Motionflow: Visual abstraction and aggregation of sequential patterns in human motion tracking data. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 21–30.
- [37] Duen-Chian Jheng, Bill Louis Harchan, Berenika Nawoja Kostka de Szttemberg, Jen-Hao Hsu, and Min-Chun Hu. 2025. Badminton Footwork Practice via an Immersive Virtual Reality System. In *International Conference on Multimedia Modeling*. Springer, 98–104.
- [38] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).
- [39] 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*. 1–17.
- [40] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James Landay. 2024. GPTCoach: Supporting physical activity behavior change with llm-based conversational agents. *arXiv preprint arXiv:2405.06061* (2024).
- [41] Marja-Leena Juntunen. 2020. Ways to enhance embodied learning in Dalcroze-inspired music education. *International Journal of Music in Early Childhood* 15, 1 (2020), 39–59.
- [42] Jihyung Kim, Jonghyeon Ka, Yelin Lee, Younghak Lee, Sangyeon Park, and Wook-sung Kim. 2022. Mixed reality-based outdoor training system to improve football player performance. In *2022 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 1–3.
- [43] Simon Kloiber, Volker Settgast, Christoph Schinko, Martin Weinzerl, Johannes Fritz, Tobias Schreck, and Reinhold Preiner. 2020. Immersive analysis of user motion in VR applications. *The Visual Computer* 36, 10 (2020), 1937–1949.
- [44] Simon Kloiber, Nicole Weidinger, Eva Eggeling, Reinhold Preiner, Katharina Krösl, and Tobias Schreck. 2022. Immersive analytics for ergonomics evaluation in virtual reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 430–433.
- [45] Lyndon Krause, Damian Farrow, Ross Pinder, Tim Buszard, Stephanie Kovalchik, and Machar Reid. 2019. Enhancing skill transfer in tennis using representative learning design. *Journal of Sports Sciences* 37, 22 (2019), 2560–2568.
- [46] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
- [47] Chunggi Lee, Sanghoon Kim, Dongyun Han, Hongjun Yang, Young-Woo Park, Bum Chul Kwon, and Sungahn Ko. 2020. GUIComp: A GUI design assistant with real-time, multi-faceted feedback. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

- [48] Chunggi Lee, Tica Lin, Hanspeter Pfister, and Chen Zhu-Tian. 2024. Sportify: Question Answering with Embedded Visualizations and Personified Narratives for Sports Video. *IEEE Transactions on Visualization and Computer Graphics* (2024).
- [49] Marc Leman. 2007. *Embodied music cognition and mediation technology*. MIT Press.
- [50] Amit Arnold Levy and Mor Geva. 2024. Language Models Encode Numbers Using Digit Representations in Base 10. *arXiv preprint arXiv:2410.11781* (2024).
- [51] Jianhua Lin. 2002. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (2002), 145–151.
- [52] 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*. 1–13.
- [53] Tica Lin, Chen Zhu-Tian, Johanna Beyer, Yingcai Wu, Hanspeter Pfister, and Yalong Yang. 2023. The ball is in our court: Conducting visualization research with sports experts. *IEEE Computer Graphics and Applications* 43, 1 (2023), 84–90.
- [54] Tica Lin, Chen Zhu-Tian, Yalong Yang, Daniele Chiappalupi, Johanna Beyer, and Hanspeter Pfister. 2022. The quest for omnisculars: Embedded visualization for augmenting basketball game viewing experiences. *IEEE transactions on visualization and computer graphics* 29, 1 (2022), 962–972.
- [55] Jingyuan Liu, Nazmus Saquib, Chen Zhu-Tian, 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* (2022).
- [56] Pin-Xuan Liu, Tse-Yu Pan, Hsin-Shih Lin, Hung-Kuo Chu, and Min-Chun Hu. 2023. VisionCoach: design and effectiveness study on VR vision training for basketball passing. *IEEE Transactions on Visualization and Computer Graphics* 30, 10 (2023), 6665–6677.
- [57] Inês Lobo, Janin Koch, Jennifer Renoux, Inês Batina, and Rui Prada. 2024. When Should I Lead or Follow: Understanding Initiative Levels in Human-AI Collaborative Gameplay. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2037–2056.
- [58] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.
- [59] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [60] Caitlyn Martinez, Seth Garbett, Kristen Hiromasa, Rhandi Jackson, Eric Miya, Michelle Miya, Joshua D White, Brian S Baum, and Mark F Reinking. 2022. Comparison of 2-D and 3-D analysis of running kinematics and actual versus predicted running kinetics. *International Journal of Sports Physical Therapy* 17, 4 (2022), 566.
- [61] Michael Muller. 2014. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*. Springer, 25–48.
- [62] Michael J Muller and Sandra Kogan. 2012. Grounded theory method in human-computer interaction and computer-supported cooperative work. *The Human-Computer Interaction Handbook* 3 (2012).
- [63] David L Neumann, Robyn L Moffitt, Patrick R Thomas, Kylie Loveday, David P Watling, Chantal L Lombard, Simona Antonova, and Michael A Tremeeer. 2018. A systematic review of the application of interactive virtual reality to sport. *Virtual Reality* 22 (2018), 183–198.
- [64] Luc Nijs, Micheline Lesaffre, and Marc Leman. 2009. The musical instrument as a natural extension of the musician. In *Proceedings of the 5th Conference of Interdisciplinary Musicology*. LAM-Institut Jean Le Rond d'Alembert Paris, 132–133.
- [65] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [66] Kelsey O'Neill and Raymond Miltenberger. 2020. The effect of behavioral skills training on shot performance in field hockey. *Behavioral Interventions* 35, 3 (2020), 392–401.
- [67] Fabian W Otte, Sarah-Kate Millar, and Stefanie Klatt. 2019. Skill training periodization in “specialist” sports coaching—an introduction of the “PoST” framework for skill development. *Frontiers in sports and active living* 1 (2019), 61.
- [68] Keith Poitier Performance. [n. d.]. 7 best moves in basketball. <https://www.youtube.com/shorts/kUumL0ObMwg>. Accessed: 2025-04-09.
- [69] Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* 11, 5 (2007), 561–580.
- [70] Randi A Sarokoff and Peter Sturmey. 2004. The effects of behavioral skills training on staff implementation of discrete-trial teaching. *Journal of applied behavior analysis* 37, 4 (2004), 535–538.
- [71] Huang-Chia Shih. 2017. A Survey of Content-Aware Video Analysis for Sports. *IEEE TCSVT* 28, 5 (2017), 1212–1231.
- [72] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.
- [73] Alexander Skulmowski and Kate Man Xu. 2022. Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational psychology review* 34, 1 (2022), 171–196.
- [74] Rhys Tribolet, William Bradshaw Sheehan, Andrew Roman Novak, Mark Langley Watsford, and Job Fransen. 2022. How does practice change across the season? A descriptive study of the training structures and practice activities implemented by a professional Australian football team. *International Journal of Sports Science & Coaching* 17, 1 (2022), 63–72.
- [75] Wan-Lun Tsai, Ming-Fen Chung, Tse-Yu Pan, and Min-Chun Hu. 2017. Train in virtual court: Basketball tactic training via virtual reality. In *Proceedings of the 2017 ACM Workshop on Multimedia-based Educational and Knowledge Technologies for Personalized and Social Online Training*. 3–10.
- [76] Wan-Lun Tsai, Tse-Yu Pan, and Min-Chun Hu. 2020. Feasibility study on virtual reality based basketball tactic training. *IEEE Transactions on Visualization and Computer Graphics* 28, 8 (2020), 2970–2982.
- [77] Carnegie Mellon University. [n. d.]. CMU Graphics Lab Motion Capture Database. <http://mocap.cs.cmu.edu/>. Accessed: 2025-04-09.
- [78] Janet Van Der Linden, Erwin Schoonderwaldt, Jon Bird, and Rose Johnson. 2010. Musicjacket—combining motion capture and vibrotactile feedback to teach violin bowing. *IEEE Transactions on Instrumentation and Measurement* 60, 1 (2010), 104–113.
- [79] L Verburgh, EJA Scherder, PAM Van Lange, and J Oosterlaan. 2016. The key to success in elite athletes? Explicit and implicit motor learning in youth elite and non-elite soccer players. *Journal of sports sciences* 34, 18 (2016), 1782–1790.
- [80] Peng Wang. 2021. Research on sports training action recognition based on deep learning. *Scientific Programming* 2021, 1 (2021), 3396878.
- [81] Jian-Jia Weng, Calvin Ku, Jo Chien Wang, Chih-Jen Cheng, Tica Lin, Yu-An Su, Tsung-Hsun Tsai, You-Yi Lin, Lun-Wei Ku, Hung-Kuo Chu, et al. 2025. Bridging Coaching Knowledge and AI Feedback to Enhance Motor Learning in Basketball Shooting Mechanics Through a Knowledge-Based SOP Framework. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [82] Jason C Wiley, Raymond G Miltenberger, and Sharayah Tai. 2024. Behavioral skills training produces acquisition and generalization of run-blocking skills of high school football players. *Journal of Applied Behavior Analysis* 57, 4 (2024), 926–935.
- [83] Margaret Wilson. 2002. Six views of embodied cognition. *Psychonomic bulletin & review* 9, 4 (2002), 625–636.
- [84] Nicola Wood, Chris Rust, and Grace Horne. 2009. A tacit understanding: The designer’s role in capturing and passing on the skilled knowledge of master craftsmen. *International Journal of Design* 3, 3 (2009).
- [85] 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*. 1–15.
- [86] Jianwen Yin, Michael S Miller, Thomas R Ioerger, John Yen, and Richard A Volz. 2000. A knowledge-based approach for designing intelligent team training systems. In *Proceedings of the fourth international conference on Autonomous agents*. 427–434.
- [87] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence* 46, 6 (2024), 4115–4128.
- [88] Zhonghan Zhao, Wenhao Chai, Shengyu Hao, Wenhao Hu, Guan hong Wang, Shidong Cao, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. 2025. A survey of deep learning in sports applications: Perception, comprehension, and decision. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- [89] Songlin Zhu and Thomas Jenkyn. 2023. Development of a clinically useful multi-segment kinetic foot model. *Journal of foot and ankle research* 16, 1 (2023), 86.
- [90] Chen Zhu-Tian, Qisen Yang, Jiarui Shan, Tica Lin, Johanna Beyer, Haijun Xia, and Hanspeter Pfister. 2023. iball: Augmenting basketball videos with gaze-moderated embedded visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

## A Appendix

### A.1 Task 1: Questionnaires for Verbal Feedback Evaluation

#### A.1.1 *Real Coach's Feedback.*

- Look up! Hold the hang a bit longer before going into the leg-through. Open up your upper body more.
- Look up! Focusing on vertical movement of your center of gravity will improve your motion. After the leg-through, imagine pushing off with your right foot and pulling through with your left foot. (Cross Drive!)
- Look up from a lowered head position. Try shifting from a smooth motion to a more rhythmic movement. Keep your upper body more upright while moving.
- Try to look up with a wider base. Focusing on vertical movement of your center of gravity will improve your motion.
- Look up. Keep your upper body more upright while moving. Try adding an eye fake when doing the in-and-out move.

#### A.1.2 *Verbal Feedback from Baseline.*

- You're moving your left elbow and knee too much to the side. Also, there's excessive movement in the right knee and ankle. Try to limit the sideways movement of your left arm and focus on more controlled, straight movements with your right leg for better balance.
- Your leg movements, particularly with your knees and ankles, are too excessive. It's particularly notable in the right movements of your right knee and left movements of your right ankle. For the upper body, there are frequent excessive movements with your left shoulder, specifically towards the left and downwards. Your left elbow too shows similar trends. You can improve by limiting these movements - try to be more controlled and precise, especially when moving to the right with your knee and to the left with your shoulder and elbow.
- Your leg movements seem to be too much and in the wrong direction. This might be causing bad balance and coordination. Try focusing on moving your knees a bit more to the right and stop moving downwards so much. This change should also help your ankles move properly. On the upper body, your right shoulder and elbow are also moving excessively to the left and down. Pay attention to limiting this movement at a faster pace.
- You're moving your arms, especially your elbows, in too many directions. Additionally, your legs, particularly the knees, seem to shift excessively. Try focusing on limiting sideways movement of your elbows and up-down motion for your knees to maintain better balance and coordination.
- Your movements are quite excessive in both your legs and arms. For your legs, specifically the knees and ankles, try to limit your movement to the sides and in the up and down directions. In terms of your arms, focus on your shoulders, elbows, and wrists. Try to also limit their movement sideways, up and down, and backwards. Less is more in this case - smaller, controlled movements can significantly improve your coordination and balance.

#### A.1.3 *Verbal Feedback from ViSTAR.*

- It looks like the main issue is with your left leg, particularly your foot. You're moving it too much to the left. Try to balance your movement by limiting the motion towards the left. Additionally, ease up your left shoulder movement to the left for a more coordinated movement at a faster pace.
- You're overextending your left arm and leaning too far with your upper body. Try to limit the forward movement of your left wrist and the rightward shift of your body. Remember, keep the motion controlled and precise for a better balance.
- It seems like you're having issues with moving your left foot and left shoulder too much to the left. Try to limit your motion in that direction for better balance and control. Don't forget, it's important to move smoothly and not overextend any particular movement at a faster pace.
- Your left leg is moving excessively upward and to the right, try to reduce these movements. Similarly, moderate the movement of your left arm upwards and to the right to maintain a steady posture.
- Your left leg movements are too pronounced - try refining them particularly in the right and downward direction. Similarly, ease movement of your left arm to the right, upwards and backwards for better balance.