Video2MR: Automatically Generating Mixed Reality 3D Instructions by Augmenting Extracted Motion from 2D Videos

Keiichi Ihara
University of Tsukuba
Tsukuba, Japan
kihara@iplab.cs.tsukuba.ac.jp

Kyzyl Monteiro
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kmonteir@andrew.cmu.edu

Mehrad Faridan
University of Calgary
Calgary, Alberta, Canada
MaKami College
Calgary, Alberta, Canada
mehrad.faridan1@ucalgary.ca

Rubaiat Habib Kazi
Adobe Research
Seattle, Washington, USA
rubaiat.habib@gmail.com

Ryo Suzuki
University of Colorado Boulder
Boulder, Colorado, USA
ryo.suzuki@colorado.edu

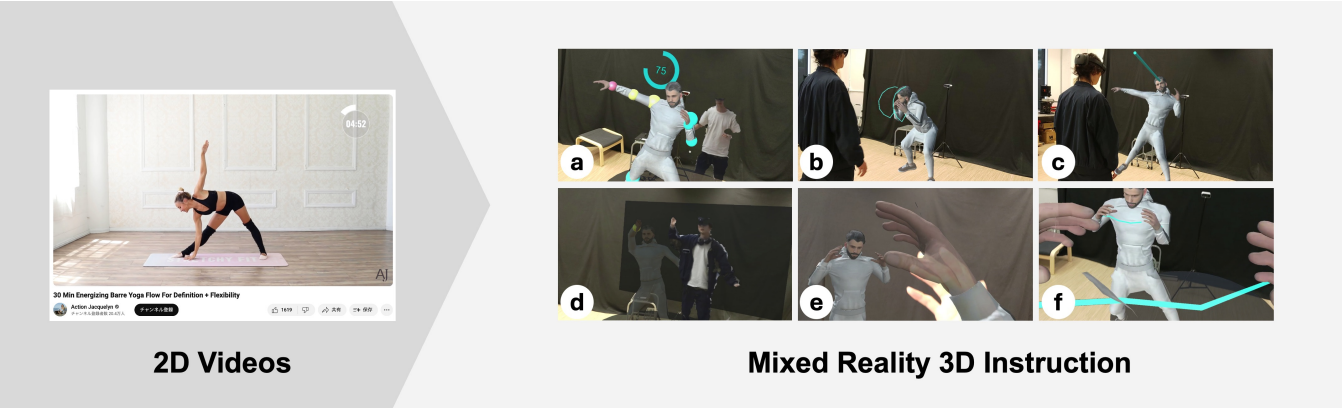

Figure 1: We introduce Video2MR, an enhanced mixed reality instruction that extract body motion from 2D videos. Video2MR augments the instructions by (a) comparing the user's movements with the instructor's movement, (b, c) visualizing the instructor's avatar by showing trajectories and highlighting gaze, (d) navigating the avatar motion based on the user's movement, (e, f) repositioning the avatar in first-person and highlighting them.

Abstract

This paper introduces Video2MR, a mixed reality system that automatically generates 3D sports and exercise instructions from 2D videos. Mixed reality instructions have great potential for physical training, but existing works require substantial time and cost to create these 3D experiences. Video2MR overcomes this limitation by transforming arbitrary instructional videos available online into MR 3D avatars with AI-enabled motion capture (*DeepMotion*). Then, it automatically enhances the avatar motion through the following augmentation techniques: 1) contrasting and highlighting differences between the user and avatar postures, 2) visualizing key trajectories and movements of specific body parts, 3) manipulation of time and speed using body motion, and 4) spatially repositioning avatars for different perspectives. Developed on Hololens 2 and Azure Kinect, we showcase various use cases, including yoga, dancing, soccer, tennis, and other physical exercises. The study results confirm that Video2MR provides more engaging and playful learning experiences, compared to existing 2D video instructions.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**.

Keywords

Mixed Reality; Sports and Exercises; Videos; Motion Capture; Avatar; Automated Generation

ACM Reference Format:
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 *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025,*

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712159

Cagliari, Italy. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3708359.3712159

1 Introduction

Mixed reality instructions have great potential for sports and exercise training. They offer an immersive and interactive learning experience that are not possible with traditional 2D video instructions. For example, prior works have shown that the ability to see a 3D avatar can improve their understanding of the postures [16, 25] and interactive MR visualizations are effective guidance to learn and notice differences in contrast to merely imitating actions from a 2D screen [48].

However, creating high-quality mixed reality instructions often requires substantial time and cost. The process typically requires 3D motion capture and programming, which presents a significant technical challenge for creating 3D instructions for professional instructors. In addition, as the manual creation process is time-consuming and tedious, these challenges significantly limit the scalability, availability, and diversity of mixed reality 3D instructions.

In this paper, we explore the idea of *automatically generating mixed reality 3D instructions* by transforming existing online 2D videos into immersive sports and exercise training. Our idea is driven by advances in computer vision and generative AI [13, 58], as well as the recent democratisation of these techniques [9], which enables us to extract 3D human motions from arbitrary 2D videos. With this, we can leverage a vast variety of professional videos already available online (eg. YouTube) to create immersive instructional experiences for various physical activities, such as yoga, dancing, exercise, and many other sports. However, several important questions remain: 1) What are the design challenges and limitations of automatically generated 3D instructions from the user's perspective? 2) How could we design and improve an automated 3D avatar instructor for a better training experiences? 3) What is the effectiveness of automatically generated Mixed Reality 3D instructions, compared to 2D videos? 4) Can this automated approach generate effective instructions across different kinds of sports and physical activities?

This paper investigates these research questions by adapting *Buchenau's 2-step experience prototyping approach* [5]. To gain insights from the user's experiences, we first develop a simple working prototype that extracts human motions from 2D videos *DeepMotion* [9], translates them to a 3D avatar and presents it in a MR spatial experience via a Hololens 2. We evaluated this minimal prototype through a formative study with 8 participants to identify benefits and challenges of automated 3D instructions. Participants found that mixed reality instructions have significant benefits as compared to their previous experiences with 2D online tutorial videos, even when automatically generated. For example, the participants expressed that they felt the increased co-presence of the instructor and appreciated the ability to see the instructor from different angles. On the other hand, areas for improvement included - difficulty in comparing their movements to the instructor's, tracking specific body parts, navigating time and controlling speed, and switching between different viewing perspectives.

To address these challenges, we present Video2MR, a system that leverages extracted human motion from existing videos, to generate a 3D avatar. Video2MR then further augments and enhances this avatar to create a mixed reality instructional experience. Our concept builds upon the previous research in MR instructional visualisations[21, 23, 63] and body-based experiences which utilize 3D avatars [16, 20, 25], but we make two key contributions beyond them.

First, we explore a **broader design space** of the enhancements and 3D augmentations of a 3D avatar in an immersive AR instructional experience. Based on the formative study and informed by previous literature, we identify our design space which includes: 1) **Posture Comparison:** contrasting and highlighting differences between the user's and instructor's avatar postures, 2) **Motion Visualization:** visualizing key trajectories and movements of specific body parts, 3) **Embodied Temporal Navigation:** manipulation of time and speed using body motion, and 4) **Avatar Repositioning:** spatially repositioning avatars to view the avatar from different perspectives. These four features explore augmentation techniques like the color indicator and scoring for comparison, footprints and trajectories for visualization, body motion-driven temporal navigation, and switching between the first-person view and third-person view. These design space elements have been inspired by prior work, for example *LightGuide* [44] for indicators, *RealitySketch* [45] for pose match, *ARrow* [23] for trajectories, *Projection based AR* [41] for footprints, *ReactiveVideo* [7] for body-based temporal navigation, *OneBody* [21] for first-person, and *OutsideMe* [59] for third-person view. While these have been introduced individually in previous works, in our design space, we aggregate and combine these to apply them to a new application scenario, an immersive 3D AR instructional experience.

Second, we contribute to a **holistic user evaluation** of the system to better understand the usability and the versatility of the system. To this end, we design and conduct a study with three parts which include: 1) a usability study with 12 participants that compares Video, Avatar, and Video+Avatar 2) a versatility evaluation which checks the accuracy and feasibility of the system across six physical activities and 3) expert interviews with five domain experts to gain in-depth qualitative feedback about Video2MR. The usability study results confirm that our approach helps automatically generate MR instructional experiences which are generally more engaging, fun, and easy-to-follow as compared to video tutorials. The expert review confirms the value of our system and our features. Also, they provided valuable feedback on each feature's practical uses, unique qualitative insights of potential use cases, educational benefits, and direction for future feature improvements. Finally, we discuss the limitations of our approach and explore future opportunities for automated MR instructional experiences.

Finally, this paper contributes:

(1) Insights from eight users through an experience prototyping protocol that elicits potential benefits and challenges with automatically generated mixed reality instructions and a design space of features that cater to these needs.
(2) Video2MR, a system that automatically generates 3D avatar animations and augmentations from 2D videos in a mixed reality immersive setting.

(3) Results and insights gained from an evaluation study with three parts which evaluates the Video2MR system and design space quantitatively and qualitatively.

2 Related Work

In this research, we developed a system that automatically generates instructional experiences in MR by creating 3D avatars from videos and automatically augmenting them with several features. This section presents prior research on the use of 3D avatars and the variety of features proposed in the previous instructional systems.

2.1 Usage of 3D Avatars

Systems employing 3D avatars have been proposed in the context of remote instruction and collaboration. These avatars can be categorized into three types: avatars that reflect real-time movements of the remote users, pre-designed avatars, and automatically generated avatars.

Remote Manipulated Avatars: There are several studies utilizing motion capture of experts in remote locations. For instance, *OneBody* [21] presents the posture of a remote person from a first-person perspective. In the context of remote collaboration, there are studies displaying the full body of a remote person [24, 49], parts of the body (e.g., hands [2, 14], gaze [4]), avatars of different sizes [37, 38], and multiple avatars [47]. These enable detailed feedback in real time. However, these systems require the presence of an expert, making it difficult for many users to utilize them easily.

Pre-designed Movement Avatars: There is also research that utilizes avatars with pre-designed movements [6, 20, 25–27, 36, 55]. For example, *ARenhanced Workout* [55] combines 3D avatars with visualizations to facilitate users' understanding of correct posture. *My Chi Coaches* [20] use multiple 3D avatars to offer views from various angles. Furthermore, in the context of tutorial creation, there are studies capturing the movements of an expert to be replayed later as avatar movements [22]. These approaches allow for the setting of various movements and the creation of accurate avatars, but creating these movements can be time-consuming. Additionally, users are limited to using pre-made movements, unable to obtain specific instructions they want immediately.

Automatically Generated Avatars: Research also exists on the automatic generation of avatars from videos for purposes such as tutorial generation [11], motion-based browsing [18], and animation authoring [54]. These studies enable the creation of 3D interactive content distinct from 2D content. However, previous research has not utilized this approach for creating MR instructional experiences, nor has it investigated what functionalities can be added to the generated avatars and which of those functionalities are useful. Therefore, we employed AI-based tools for 3D avatar generation to create MR instructional experiences in 3D, proposed various functionalities, conducted user studies, and investigated the effectiveness of these functionalities.

2.2 Features in Instructional Systems

Various instructional experiences have been proposed in past research, not only in MR environments but also using 2D screens, projections, and VR devices. Here, we show the types of instructional experiences proposed in other media and contexts and how we have applied them in our 3D MR experience.

Visualization Techniques: A wide range of visualization techniques have been proposed for instruction systems. For example, in the use of projection, several visualizations have been proposed, such as visualizing footprints [29, 41], indicating correct directions [44], using metaphors [42], and visualizing users' posture [52]. Additionally, using the 2D screens, displaying trajectories on 2D screens [23], and presenting synchronization accuracy [63] are also explored. Moreover, various 3D visualization techniques have been proposed, such as visualizing the movement of a badminton shuttle in VR [60] and visualizations for ski training in VR [35]. In AR, several visualization techniques have been proposed, such as visualization techniques for workouts with 3D avatars [55] and for improving free throws in basketball practice [31]. In contrast, our focus is on visualization applied to generated avatars. By automatically generating visualized instructional experiences from videos, we support a wide range of sports and exercises.

Comparison Methods: Many studies have proposed using stick figures on 2D screens for comparison [8, 10, 34, 39, 48, 53, 62]. These studies compare the user's 2D posture with the instructor's posture [3, 48, 53]. For example, *YouMove* [3] and *MotionMA* [53] compared the user's input video with the instructor's movements. Tharatipyakul et al. [48] conducted a study comparing feedback using videos and stick figures. Other studies compare using 3D avatars on 2D screens [15, 32]. *AIFit* [15] converts input videos into 3D postures and provides feedback on differences from correct postures. *PoseCoach* [32] is a study that utilizes 3D avatars in 3D space. Our system compares the user's 3D posture captured by an RGBD camera with the automatically generated instructor's 3D posture.

Temporal Navigation Techniques: To change the timing in videos, mainly 2D UI elements like scroll bars are used, but different methods, such as those using avatars and body movements, have also been proposed. For example, Hamanishi et al. [18] arrange avatars in a timeline and manipulate time by interacting with them. There are also studies that compare the user's 2D posture with the 2D posture in videos to navigate to specific times [7, 17]. In contrast, our system proposes changing the generated avatar's 3D posture based on the user's 3D posture.

Utilization of First-Person Cues: Several studies use first-person cues to instruct on how to move their bodies [12, 19, 21, 61]. There are also studies specialized in specific applications, such as Aikido [46], musical instrument performance [33, 43], agility training [29], and juggling [1]. However, these first-person cues are manually created, and their production can be time-consuming. Therefore, we aim to automatically generate avatars and then convert them into first-person cues.

Multimodal Feedback: Instructional methods utilizing not only visual but also auditory and haptic feedback have been proposed [28, 40, 56]. For example, *VoLearn* [56] provides feedback on the user's movements using wearable devices and auditory cues. *Multimodal motion guidance* [40] uses vibrations to give instructions on speed and direction. Furthermore, *Stylo and Handifact* [28] offer the sensation of the hand being pressed to assist in improving

movements. In contrast, our system focuses on generating visual guidance and feedback.

3 Experience Prototyping

To identify the challenges and potential benefits of automatically generated 3D instructions, we adapted *Buchenau's experience prototyping approach* [5] by developing an initial prototype and then conducting a formative study. This is because MR experiences are often difficult to imagine before experiencing them, thus it is difficult to get an appropriate insight or design guidelines without a functional prototype. Therefore, we developed a simple prototype to test our concept and gain user feedback through the formative user evaluation.

3.1 Initial Prototype

For the initial prototype, we use off-the-shelf software that automatically extracts human motion from online 2D videos. We initially tested four different software, including *DeepMotion* [1], *Plask Motion* [2], *Kinetix* [3], and *Rokoko* [4]. After the initial investigation, we decide to use *DeepMotion* as it can produce the highest quality and most accurate results for our purpose. We converted six 2D video tutorials into 3D avatars using *DeepMotion*. These videos include tennis, dance, baseball, yoga, taichi, and exercise. We generated 30-second instructions for each video. Based on the extracted human motion, we developed a simple Unity application that shows a 3D avatar animated based on extracted body motion and the associated video on the background in the mixed reality scene through Hololens 2 (Fig 2).

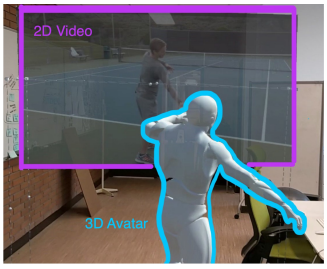

Figure 2: First-person view of the experience in the formative study

3.2 Formative Study Protocol

To better understand the benefits and challenges of using 3D avatars generated from 2D videos, we conducted a formative study with eight participants (5 males, 3 females, ages 21 - 38). All participants had previously learned physical sports like dance, soccer, yoga, martial arts, and table tennis through online videos.

Each interview for the formative study was 50-70 minutes long. All interviews were recorded and later transcribed with the consent of the participants. The purpose of the formative study was to understand the challenges users faced in 3D Mixed Reality tutorials and elicit improvements. First, participants were given a brief description of the system they would be experiencing and were then given a demo of six instructional MR experiences. The demos were aimed at giving participants an understanding of MR instructions. Following the demos, participants shared their experience of using the prototype and compared it with their previous experience in online learning. Their transcribed responses were thematically coded by two authors, the sections below highlight the themes that emerged in the form of benefits and challenges experienced and expressed by participants.

3.3 Benefits

Overall, participants appreciated the 3D avatar instruction. A significant advantage highlighted by the participants was that it was easy to understand the spatial orientation of the instructor's poses, a challenge often faced with traditional 2D formats. *"I think the advantage is clear: if I just watched a 2D video, I might not clearly understand some 3D aspects, like which leg is in front and which is at the back."* (P1). The ability to adjust their viewpoint around the 3D avatar was particularly appreciated (P3-4,P6, P8). *"I think one useful feature is being able to walk around it. Since it's 3D, it helps because you sometimes lose sight of details from certain camera angles."* (P4) *"It's nice to see the backside because sometimes that's easier to follow [...] I like that as an option and that's something that 2D videos lack"* (P3).

Another benefit that participants mentioned is the presence of the instructor in the MR environment. One participant mentioned that *"the biggest difference that I can feel is that the [avatar] model is really here I feel myself more immersed"* (P6). Another participant echoed this aspect by saying *"It feels closer to having another person explaining, rather than just watching a 2D video."* (P7).

Conversely, participants also appreciated that the instructor wasn't physically present, allowing for pressure-free learning (P5, P7). *"One thing I like about learning through visual media is that I won't be judged by another person [...] if I dance poorly I don't want others to see me dance but if I dance in front of a virtual instructor I won't be judged ... it mitigates social awkwardness"* (P5)

3.4 Challenges and Needs

Participants shared several challenges and needs they experienced and discuss how the MR experience can be improved by resolving these problems. Through the interview, we identify the following four main challenges of the current simple prototype.

3.4.1 ***Posture Comparison: Needs to Easily Compare between User's and Instructor's Motion.*** One of the benefits of mixed reality instructions highlighted by participants is that the user can see the instructors in the real-world scale (P3, P6-8). This helps participants to mimic the instructor's movements, but they felt that the current system could do more to further enhance this benefit. *"Whenever you are practicing, the avatar can only show you what to do, but it cannot tell you what you are doing wrong."* (P6) *"It would be great if you could compare your position to the 3D avatar model and if there is some way that it can detect something like - your arm isn't high enough because at least for me it's hard to know if I am doing the same thing as the video"* (P3). To address this gap,

[1] https://www.deepmotion.com/
[2] https://plask.ai/
[3] https://www.kinetix.tech/
[4] https://www.rokoko.com/

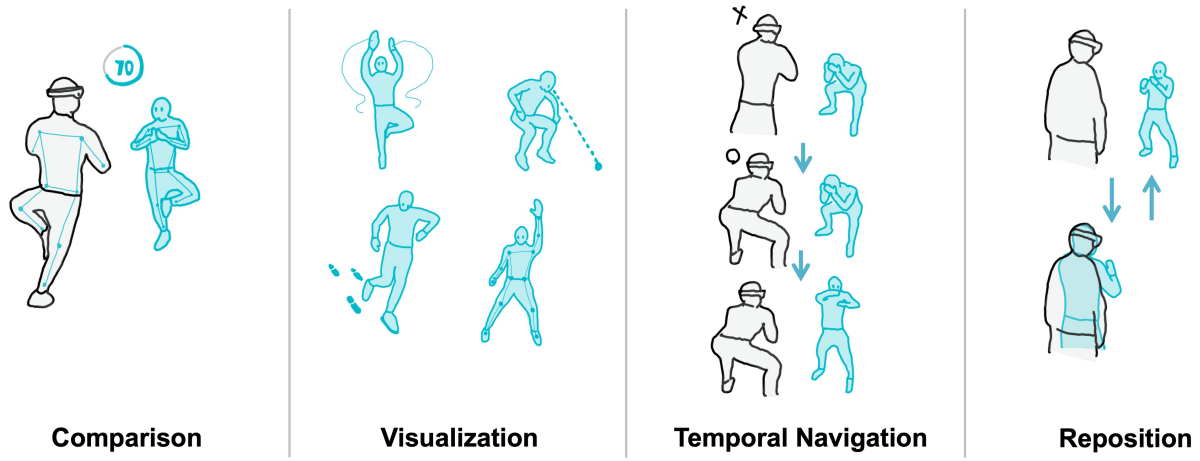

Figure 3: Design Space of Video2MR

participants suggested enhancements to the MR instructions by providing real-time feedback through comparison visualizations. *"If there were some intelligence that could analyze your motion and provide feedback like, 'You're doing this wrong; why don't you modify it?' and then show what your movement looks like compared to the correct one while adapting in real-time, that would be really cool."* (P2)

3.4.2 ***Focusing and Highlighting: Difficulty in Tracking Specific Body Parts.*** Participants also express the need to focus on the motion of specific body parts (P1-2, P4-7). *"Like when he's swinging, you want to see the hand and when he is waiting, you want to see the leg."* (P1) *"And in table tennis the ball is important but even the hand motion is important."* (P2) Participants highlighted the importance of focusing on the motion of a body part which is the point-of-interest in the instructional experience (P4, P6-7). *"I have noticed that some videos are focused on a particular body part, like for moonwalk we focus on the feet"*(P4) They also further shared visualizations which could be potentially useful. *"I was thinking of showing trails if I wanted to track one of his hands showing that it moved from here to here - showing the path of hands or feet, like in animation, you have onion skinning"* (P4)

3.4.3 ***Temporal Navigation: Challenges in Controlling Speed and Time of the Avatar Instruction.*** Participants mentioned that they find it challenging to navigate through the instructions, especially when they want to simultaneously follow the instructions and navigate through the video content (P1, P3, P5-7). They shared a strong need for an intuitive way to navigate through the instructional content (P1, P3, P5, P8). *"I usually have to stop the video when I miss a certain part of the video, sometimes rewind the part I want to mimic"* (P3). Participants also suggested techniques to make it more useful. *"If the video could adapt to my progress that would be cool because I feel sometimes I couldn't follow the videos if they are too fast (...) So if they could adapt like if I am still doing the last [step] it could slow down a little bit"* (P5). *"If I want to skip to the other pose I know, but want to re-watch, I can just do that pose. I want to control the progress of the video in a natural way, like moving your body."* (P1).

3.4.4 ***Spatial Reposition: Needs of Seamlessly Switching between First- and Third-Person Perspectives.*** Participants also mentioned they often would want to teleport to the instructor's first-person perspective, as they believe it would facilitate a more accurate replication of the instructor's actions (P2, P6). *"Sometimes with these types of instructions, we get left and right mixed up so being able to switch to a first-person view could be helpful."* (P4) Participants also mentioned that they need to alternate between watching a video and monitoring their own actions to ensure they are following instructions accurately (P2, P8). The first-person perspective could also solve this juggling problem. "that's always the challenge, you're doing a move and they're saying now do this and you're like, well, I can't look at you because you're over here and my face is turned the other way." (P8) *"...then you can't see the video, but it doesn't have to be because now that it's in [first-person] AR you can always look at it."* (P2)

4 Video2MR: Augmenting Auto Generated MR Instructions

Based on the insights, we designed Video2MR, a system that augments the automatically generated mixed reality instruction. Similar to the initial prototype, we used *DeepMotion* to convert 2D video into 3D animation, and show the avatar model chosen from Mixamo [5]. We utilized Unity (version 2020.3.35f1) and Mixed Reality Toolkit (version 2.8.3) to create our system. The scene is rendered in HoloLens2 [6] that is connected to a laptop PC (G-Tune, Intel Core i7-11800H 2.30GHz CPU, NVIDIA GeForce RTX 3060 GPU, 64GB RAM) to show the MR scene through the Holographic Remoting Player [7]. We also capture the user's body using one Azure Kinect

[5]https://www.mixamo.com/
[6]https://www.microsoft.com/en-us/hololens/
[7]https://learn.microsoft.com/en-us/windows/mixed-reality/develop/native/holographic-remoting-player

RGB-D camera [8] placed in front of the user with a tripod. The Kinect camera is connected to the same laptop through a USB-C cable. We use both skeleton position data and 3D mesh data for real-time capturing. Additionally, we utilized the Azure Kinect Examples for Unity package [9] to generate 3D meshes from RGBD data, enabling users to view their 3D posture. Azure Kinect's Body Tracking feature allowed for the estimation of joint angles, capturing 32 joint angle data from users [10]. Similarly, DeepMotion could acquire rotational data for 20 joints within a video [11]. Our system does not store or share the motion data. All motion data is used solely for on-the-fly mesh rendering and feedback within the local system, ensuring user privacy and data security.

Based on the formative study, we design the following four features: 1) ***Posture Comparison:*** contrasting and highlighting differences between user and avatar postures, 2) ***Motion Visualization:*** visualizing key trajectories and movements of specific body parts, 3) ***Embodied Temporal Navigation:*** embodied manipulation for time and speed through body motion, and 4) ***Avatar Repositioning:*** spatially repositioning avatars for different perspectives. All of these features are automatically generated based on the extracted body motion and user posture captured by Azure Kinect.

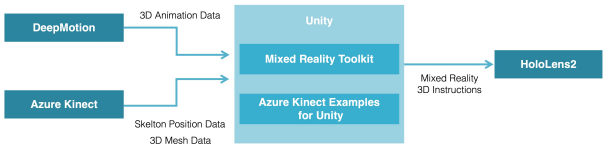

Figure 4: System Configuration

4.1 Posture Comparison

In the formative study, users sometimes wanted to learn by mimicking the instructor's movement. To support this, the system shows the user's real-time mesh on the side or top of the instructor's avatar. We not only show their mesh but also calculate the difference between the user's and instructor's posture and provide feedback by changing the color or providing the synchronization score.

4.1.1 Pose Match Indicator. The accuracy of user postures relative to the instructor's avatar is crucial for ensuring effective practice and understanding of movements. This indicator provides real-time feedback to users by visualizing the alignment discrepancies between the user's limbs and those of the avatar. Specifically, colored spheres are dynamically displayed on the avatar's limbs, including left and right arms, and left and right legs. The color of each sphere indicates the degree of alignment. Blue spheres signify close alignment, yellow indicates minor deviations, and red marks significant misalignments (Fig. 5). For example, during a complex dance sequence, users may struggle to synchronize their footwork with the instructor's. The Pose Match Indicator visually alerts them of their inaccuracies, allowing for immediate correction and improvement.

[8]https://azure.microsoft.com/en-ca/products/kinect-dk
[9]https://assetstore.unity.com/packages/tools/149700
[10]https://learn.microsoft.com/en-us/azure/kinect-dk/body-joints@:
[11]https://blog.deepmotion.com/2020/11/19/animate-3d-custom-characters/

Figure 5: Pose Match Indicator: Avatar joints turn blue when the user's movements are correct, and red when incorrect.

4.1.2 Pose Match Score. The pose match score quantifies the alignment between the user's posture and the instructor's ideal posture in an abstract way (Fig. 6). The score is shown as a numeric value and a circle graph. The score is calculated based on the difference between the instructor's and the user's body positions. We first calculate the score for each joint, add all the scores for each joint, and then show the total score. In this system, we used ten joints, each with a maximum of 10 points, and the maximum total score is 100. This could be used when overall body alignment is more important than the pinpoint accuracy of individual joints. For example, in a boxing workout, we don't have to match the posture with the instructor accurately, but we have to make the whole body posture similar to the instructor's for an extended period.

Figure 6: Pose Match Score: The score displayed behind the avatar changes based on the accuracy of the user's movements.

4.2 Motion Visualization

The user also sometimes wants to focus on the motion of specific body motions. For example, for dancing, the user wanted to focus on the foot motion. To address this, we highlight the specific motions including head gaze, footprints, and trajectory.

4.2.1 Head Gaze. Some previous works visualized the body direction by using lights [51, 52]. Inspired by these works, we visualize the head gaze of the instructor. The instructor's head gaze can provide valuable insights into where their focus is directed during the activity. Showing this can improve the users' understanding of the instructor's intent or the adequate head direction. The gaze is shown using the ray, which starts from the instructor's avatar's head (Fig. 7). For instance, in dance instruction, although it is difficult to understand the head direction while overlaying the body to the instructor's avatar, by using this, the user can understand the direction by looking at the gaze ray.

4.2.2 Trajectory. This feature displays a motion path, providing a visual guide for users to follow. In Figure. 8, the instructor's hand trajectory is shown, which shows the hand's position in the past. This can be used for users who struggle to keep up with the instructor's pace due to the complexity or speed of the movements. For

Figure 7: Head Gaze: The ray extending from the avatar's head highlights the direction of the avatar's head gaze.

instance, in tennis instruction, users can use this mode to understand and replicate fast and complex swings that might otherwise be challenging to follow.

Figure 8: Trajectory: Displays the path to reveal the temporal movement of specific body parts.

4.2.3 Footprints. Inspired by Projection-Based AR [41] that projects footprints using a projector, we utilized footprints to highlight the 3D avatar's feet positions. Footprints provide important information about foot placement, which is often occluded by the instructor's body. The foot positions of the instructor's avatar are acquired and marked with blue on the ground to indicate the footprints (Fig. 9). A new footprint appears every few seconds and subsequently fades by making it gradually transparent. This feature is beneficial in learning activities such as baseball instruction, where understanding the correct foot positioning is important for mastering the technique of throwing a ball.

Figure 9: Footprint: Displays the position of the avatar's feet, highlighting where the user should place there feet.

4.3 Embodied Temporal Navigation

Inspired by previous works, such as *Reactive Video* [7] and *PoseAsQuery* [17] that implements body-based navigation for 2D videos, we extended this concept to navigate 3D avatar motions. Our system allows the user to stop the avatar motion until the user synchronizes the exact same posture. Then, the avatar motion starts moving as if the avatar moves step-by-step, based on the user's motion. We achieve this by calculating using the scoring system mentioned in 4.1.2. We set a threshold score, and if the user's score is better than the score, the avatar will jump the animation timeline to move to the next step. We set the threshold score high if the high accuracy is important, and we set it low if it is not. Our system changes how much we jump based on whether the user wants to understand a long duration of instruction quickly or understand the detailed movement of a short duration.

Figure 10: Body-based Navigation: The avatar's posture adjusts based on the user's movements.

4.4 Avatar Repositioning

In the formative study, the users wanted to mimic their behaviors from different perspectives. Inspired by OneBody[21] and AR-Arm [19], our system allows the user to transport from the third-person view to the first-person view, and vice versa (Fig. 11). We enabled this by automatically synchronizing the user's head position to the avatar's so that the first person will be shown correctly whether the user moves around. By using the first-person view, users can just move their body parts to the first-person instructor's body part to understand the proper position of those. Also, this could be combined with visualization, including trajectory, head gaze, and footprint (Fig. 12). For example, hand trajectory in first-person view could help users follow the hand movement when the movement is rapid and expansive.

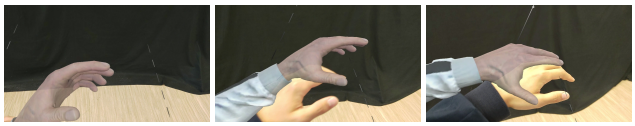

Figure 11: First-Person View: Allows the user to observe the avatar's movements from a first-person perspective.

Figure 12: First-Person Visualization. Trajectory (left), Head Gaze (center), and Footprint (right).

4.5 Versatility Evaluation

To determine what kinds of online sports videos are and are not suitable for our system, we conducted a versatility evaluation.

4.5.1 The Evaluation Criteria and Dataset. We selected 10 online sports videos from diverse sports categories: Yoga [12], Dancing [13], Martial arts [14], Gym workout [15], At-home workout [16], Swimming [17], Baseball [18], Tennis [19], Boxing [20] Fencing [21]. To avoid cherrypicking

[12] https://www.youtube.com/watch?v=j7rKKpwdXNE
[13] https://www.youtube.com/watch?v=cUJRn-WfTbw
[14] https://www.youtube.com/watch?v=e64AtWekQVo&t=395s
[15] https://www.youtube.com/watch?v=sthD8ziGP1c
[16] https://www.youtube.com/watch?v=cbKkB3POqaY
[17] https://www.youtube.com/watch?v=LijdyVaaDnY&t=2s
[18] https://www.youtube.com/watch?v=YY9tErIBVQw
[19] https://www.youtube.com/watch?v=CXgfNBnetzQ&t=341s
[20] https://www.youtube.com/watch?v=kKDHdsVN0b8&t=170s
[21] https://www.youtube.com/watch?v=s-bXSFtXqQk&t=1020s

videos that would work well for our system, we selected videos popular on YouTube, at least within the top 5 videos. Furthermore, to conduct our evaluation in a standardized manner, we searched for the online videos by using the same search prompt: "Category Name" + "Tutorial" or "Lesson" on YouTube. We chose videos between 3 and 5 minutes long, and for longer videos, we trimmed to 5 minutes that featured the instruction's main performance.

We measured the following metrics: Success Minutes, Success Rate, and Number of Errors. Success is defined as the total time during which the system accurately tracked and visualized the instructor's movements. Errors refer to instances where tracking failed, such as when body parts were occluded by clothing, motion was too fast, or the system temporarily lost accuracy. Sometimes causing irregular or impossible movements by the avatar or extreme distortion of it's linbs. Number of Errors indicates how many times tracking failures occurred with their total duration noted in parentheses for clarity (e.g., 10 errors over 32 seconds). Success Minutes refers to the cumulative time when the system performed successfully, and the Success Rate describes as a percentage how much of the total session time was accurately tracked. We also included certain factors that were known to influence the system's output quality, including clothing occlusion, speed of motion, different lighting conditions and the placement of the camera relative to the person being filmed.

	# of Camera Angles	Total Minutes	Success Minutes	Success Rate	Num of Errors	Camera Placement	Character Placement	Lighting	Occlusion	Clothing	Speed of Body Motion
Yoga	2	5:00	4:28	89%	10 (32s)	Stationary	Center	Good	None	Good	Slow
Dancing	1	5:00	4:45	95%	5 (15s)	Stationary	Center	Good	Self	Baggy	Various
Martial Arts	1	5:00	4:19	86%	11 (1:41)	Stationary	Center	Good	Self	Baggy	Various
Gym Workout	4	3:09	1:44	55%	18 (1:25)	N/A	N/A	N/A	Equiptment	Good	Slow
At-home Workout	2	5:00	4:02	81%	14 (58s)	Stationary	Center	Good	Self	Good	Fast
Swimming	3	5:00	0:03	0%	Infinite	Moving	Center	N/A	N/A	Good	Fast
Baseball	2	5:00	4:38	93%	7 (21s)	Stationary	Center	Good	Equiptment	Good	Various
Tennis	1	5:00	4:32	91%	8 (28s)	Stationary	Center	Good	Equiptment	Good	Medium
Boxing	1	4:48	4:23	91%	5 (25s)	Stationary	Mostly Center	Good	None	Yes but Negligible	Medium
Fencing	1	4:43	4:27	94%	6 (16s)	Moving	Moving around	Good	None	Yes but Negligible	Fast

Figure 13: Versatility Evaluation Results

4.5.2 Results and Findings. The table above shows the result of our evaluation. Most videos performed quite well, above 85% accuracy, though all videos had visual glitches to some level. Swimming performed the worst by far, glitching out completely 3 seconds into the video. For most videos, the fingers were incorrect. For example, in tennis, boxing, tennis and baseball the fingers of the avatar were outstretched. However, the input video shows an instructor gripping some equipment, making a fist or otherwise not what the avatar's fingers look like. Additionally, if the input video had the instructors hands close together, then it's much more likely that the avatar's arms would clip through the other arm or the torso. For yoga, at-home workout and especially martial arts, tracking suffered when the instructor was on the floor. In particular for martial arts, the baggy and mono-colored martial arts clothing made tracking accuracy very poor at times. The gym workout's lower accuracy was primarily due to occlusion issues as well as numerous camera angles.

5 User Study

To evaluate the effectiveness of Video2MR, we conducted a user study comparing Video2MR with 2D videos. Also, we evaluated the usefulness for each feature for six videos.

5.1 Method

5.1.1 Participants. We recruited 12 participants (7 male, 5 female), aged between 21-38 years ($M = 25.17, SD = 4.73$). To evaluate their familiarity with video instructions, we surveyed it using a 7-point Likert scale, ranging from 1 (Not familiar at all) to 7 (Extremely Familiar). The mean score of the familiarity was 6.42 ($SD = 0.79$).

5.1.2 Conditions. We evaluated the system under the following three conditions:

Video : Participants viewed the videos on an iPad, using the default video player.

Avatar : Participants viewed a MR scene using Hololens2, showing avatars both the instructor's avatar and the user's mesh positioned in front of the user.

Avatar + Video This condition combined both avatars and video. The MR scene displayed avatars with the video running in the background through Hololens2.

We conducted the study in a with-in-subjects design. Therefore, each participant used three conditions.

5.1.3 Study Setup. Figure. 14 shows the setting of our study. Hololens2 was connected to a laptop PC through the Holographic Remoting Player. Also, the Azure Kinect camera was connected to the laptop PC though a USB-C cable. The experimenter controlled which feature to display using Unity. The video and avatar size and position were customized based on individual participant preferences. We selected videos from six various categories: yoga, dance, martial arts, tennis, soccer, and exercise (Fig. 15). Each video was one minute long.

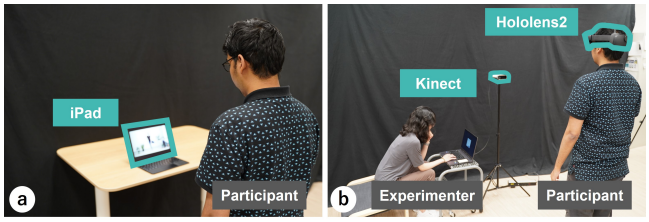

Figure 14: Study Setup. (a) *Video* condition (b) *Avatar* and *Avatar+Video* condition.

5.2 Study Design

5.2.1 Procedure. First, we asked the participants to provide their consent. Then, we conducted the pre-study questionnaire to ask about their familiarity with video instructions. The study consisted of six sessions. In each of these sessions, participants viewed one

of the six different videos, each under the three conditions. We told the participants to follow the instructor's movement and move their bodies to learn the correct movement. To counterbalance the order effect of the conditions, we used six different orders of the three conditions for six sessions. Also, to counterbalance the order effect of the videos, we used six different orders of the six videos using a Latin square design. Therefore, one video order was used by two participants. While we were conducting *Avatar* and *Avatar+Video*, we showed our system's features for each session. After each session, they answered how useful each feature was. After all sessions, participants answered the questionnaire about their overall experience with each condition. Finally, we interviewed the participants to gather feedback. The study was approximately 75 minutes. They were compensated 15 USD (The actual currency has been removed for anonymity).

5.2.2 Measurements. To compare the overall experience between the three conditions, we asked the following questions using a 7-point Likert scale (1: Not at all, 7: Extremely): 1) **Co-presence**: "How much did you feel the sense of the instructor's presence?", 2) **Engagement**: "How much did you feel engaged in the instruction?", 3) **Fun**: "How fun was it to use this system?", 4) **Easy to Follow**: "How easy was it to follow the instructor's movement?". Additionally, we asked about the usefulness of the features for each video using a 7-point Likert scale.

5.3 Results

5.3.1 Overall Experienece. The results of the overall experience are shown in Fig.16. For **Co-presence**, participants rated *Avatar* and *Avatar+Video* better than *Video* (*Video*: 5.3, *Avatar*: 5.9, *Avatar+Video*: 5.8). Showing a 3D avatar and making it move with the participants might have increased the sense of the avatar being near (P3, P10). *"It made me feel that an instructor was actually standing in front of me."*(P3). *"I can see the instructor moving with me, so I felt like they're right there"*(P10).

For **Engagement**, participants rated *Avatar* and *Avatar+Video* better than *Video* (*Video*: 4.7, *Avatar*: 6.0, *Avatar+Video*: 6.3). Because the instructor is in front of the user, participants might have tried to match their postures (P3, P4, P11). *"It was engaging because I was trying to match my postures quite closely with the avatar."*(P4). *"If I was just watching a video, I wasn't following it as much. But if it was the avatar doing it, I was trying to follow it, I had my concentration there."*(P11).

For **Fun**, participants rated *Avatar* and *Avatar+Video* better than *Video* (*Video*: 4.4, *Avatar*: 6.1, *Avatar+Video*: 5.6). This could be because our system provides feedback that videos alone cannot offer. *"It's more fun when you can see that you're doing things right."*(P10) Also, it was because it is a different way of learning instruction (P4, P5). *"It was fun because it was like cool and exciting and new"*(P5)

For **Easy to follow**, participants rated *Avatar* the lowest and *Avatar+Video* the highest (*Video*: 5.3, *Avatar*: 5.1, *Avatar+Video*: 5.8). The reason could be because, in some scenarios, the avatar could not represent the nuanced movement (P5, P6). *"The avatar didn't capture the nuance of the human body. So stuff like for squats was actually really hard to follow."* and *"ones where the motion was slow and nuanced, the video was actually really easy to follow."*(P6) It was *"occasionally difficult just in terms of a bit of jitter with the tracking."*(P5). Also, it was difficult for participants to follow when the instrument was important to follow the instructions (P2, P5, P10). *"When the instructor interacted with external objects, having just the avatar, there was very hard to follow."*(P2) *"It was generally easy to follow the instructions except in cases where some of the context was missing things like the, the props, like the tennis rackets or the tennis balls or the soccer balls."*(P5)

5.3.2 Feature Evaluation. The results of the feature evaluation are shown in Figure17. **Indicator** and **Scoring** were rated high compared to other features. This helped participants to do the exact same movements with the instructor. *"The indicators were the most helpful feature because you would know whether your body is in the exact position."*(P11) *"I feel it's most useful to check whether your body part is in the right position. That's why I felt indicator and scoring the most helpful."*(P10) Also, some participants mentioned some issues with the indicator. *"I also want the indicators on myself. It was hard to look at someone's body and clearly understand which part the indicator was referencing."*(P6) *"Certain joints were not as important to the instruction and so showing them wasn't as useful."*(P4)

Body-based Navigation was rated high for dancing. The movement was fast in dancing, so this feature might help them follow the instructions by stopping until the posture is correct. For the dance scenario, *"I rated the body-based navigation highly because it would be really useful to learn the dance slowly. If the instructor did difficult poses in very quick changes, it would be useful to make sure you have the correct pose."* (P8). Also, it helped the participants to understand the sequence. *"Body-based navigation is quite useful to make sure that you're doing everything in the right order."*(P7) Soccer was rated low for Body-based Navigation. This might be because the instructor used a huge space, and it was difficult to compare the body pose to it. *"It was more challenging to use indicators in the soccer scenario because the instructor was moving around and to follow him and the precise movement was harder."*(P9)

The **First-person** feature could be the most affected feature by the technical issue of Hololens2. Because of the limited field of view of Hololens2, people suffered from following the first-person instructor. *"I wasn't always able to see the hands clearly. They were only visible when he slid them across his face."*(P1) However, the concept of using the first-person view was appreciated by some participants. For the tennis scenario, *"Fore hands maybe first person view was a bit more helpful because I could really see like the hand is this like I'm trying to match my hand. I could understand how I have to position it, maybe the angle or how high I have to raise it."*(P2) For the martial arts scenario, *"it was interesting because the first person view visualized the height the instructor was getting on his kick versus my own kick better."*(P5)

For **Footprint**, participants rated Martial Arts the best. The footprint was useful for scenarios in which the foot actively moved. For martial arts, *"Given that the stance was important for the kicking and returning to the stance was important. So I really liked the footprints this time."*(P5) On the other side, for exercise, *"I didn't find the footprint useful for this activity because you don't really move your feet too much for the squad."*(P5)

For **Head-gaze**, participants rated Martial Arts the best. The head-gaze helped participants to understand where they should look. *"Head-gaze kind of helped me to know where to point my*

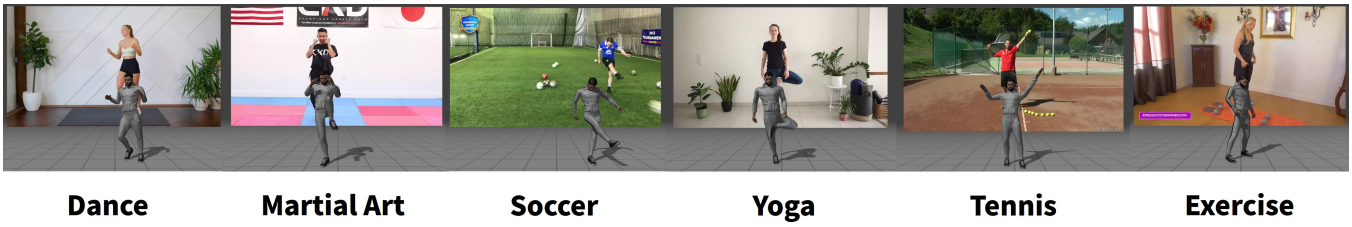

Figure 15: The videos and their categories that we used in our user study.

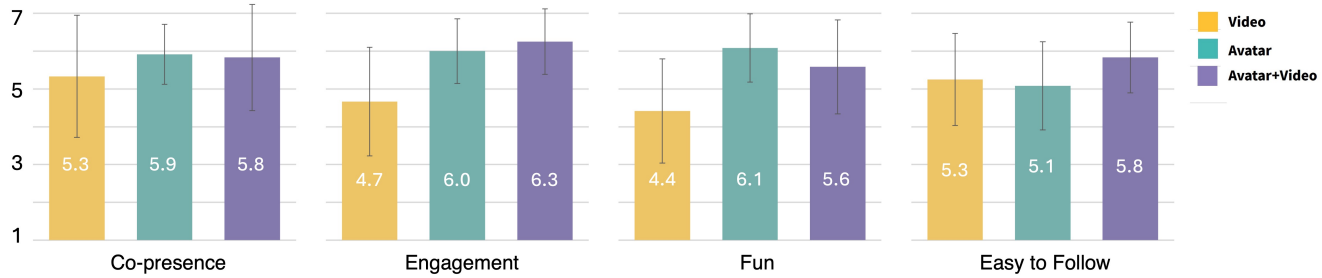

Figure 16: Results for the overall experience. The colored bars represent the mean scores and the error bars indicate the standard deviations.

head."(P4) Also, it helped how static the instructor's head is. For the yoga scenario, *"Unlike the other ones, seeing how still the instructor's gaze is really interesting. And understanding that you're not as still as them"* (P6)

For **Trajectory**, participants rated Tennis and Dance the best. This could be because the current implementation only showed the hand trajectory, so the apps that use hands often could be more useful than apps that use lower body, such as Martial Arts, Soccer, and Exercise. For the tennis scenario, *"If I had gotten more practice, hand trajectory could be useful, particularly for tennis, where your hand movements are sort of important"*(P5). For exercise, *"I needed more trajectory on the hips or the knees, more than the hands when it comes to squatting"*(P7). And the trail shape made by the hand trajectory attracted the participants. *"Seeing the arc of the hand trajectory was really interesting"*(P6).

5.3.3 Sports Category-Based Evaluation. For Dancing and Martial Arts, participants preferred *Avatar* to *Avatar+Video*. The reason could be that the avatar information was enough to follow the instructions (P2, P4). For dancing, *"I don't really see the need for the video because it's way easier to just follow the avatar because it's doing the same thing with the video."*(P2) Another participant mentioned for dancing, *"Avatar+video was just really difficult because there's a lot of information."*(P4)

Although, for Tennis, participants prefer *Avatar+Video*. For tennis, it was difficult for participants to understand what the instructor was doing by watching the avatar without the video. For tennis, just showing avatar, *"I didn't know which hand had the tennis racket and which one was throwing the ball, for instance."*(P4)

Martial Arts was the most preferred video. The reason might be that in martial arts, the 3D position of the body position was important and the avatar helped them to understand it. *"It was actually quite useful to look at the avatar from all angles. It really helps particularly with the leg positioning."*(P6)

On the other side, Exercise was the least preferred video. The reason might be the movement was too simple. *"The squatting one, the avatar was kind of unnecessary because exactly what the instructor is doing you can just do."*(P7)

6 Expert Review

6.1 Method

To conduct our expert review, we recruited six experts in the physical exercise domains. Our experts were between 25 - 36 years old, with 4 - 30 years of experience within their domain of expertise. Each participant was an expert in one of the following domains: (EY) yoga, (ED) dancing, (ES) soccer, (EM) martial arts, (EF) fitness and (EG) golf. To evaluate our system's suitability for their domain of expertise, We selected a 60-second clip from YouTube of a video tutorial of their domain, showed the experts the video clip on an iPad Pro's screen then had them experience Video2MR. We then walked through each of our implemented features, asking them various inquisitive questions to gather qualitative feedback. Overall, watching a video clip to experience our system and the subsequent interviews took about 60 minutes to complete. Expert participants were compensated 20$ USD (The actual currency has been removed for anonymity).

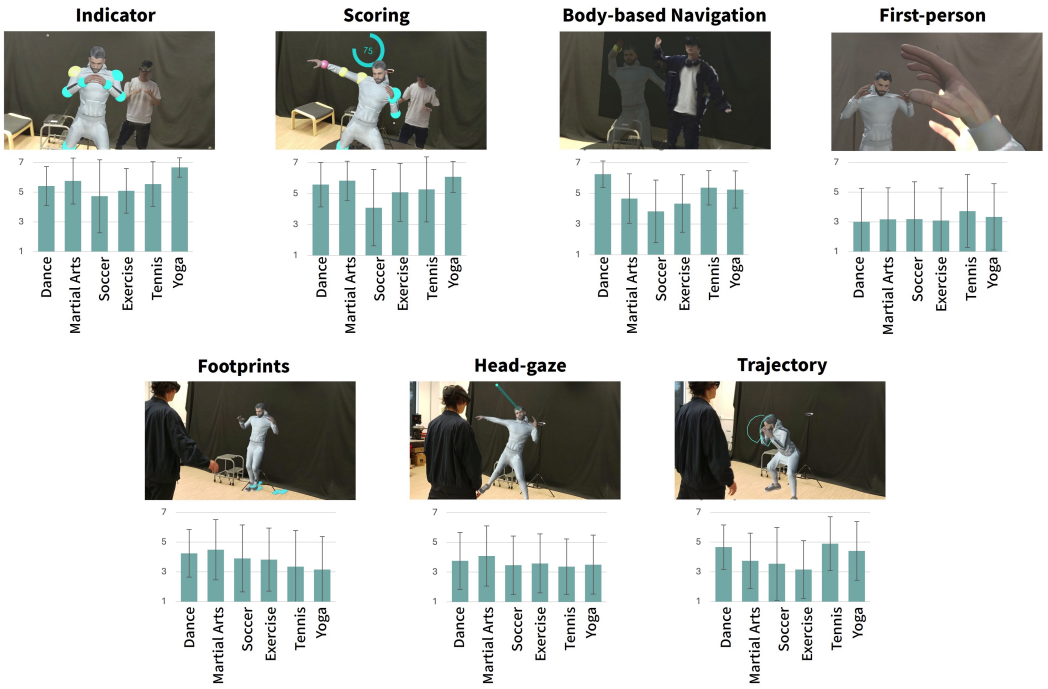

Figure 17: Results of feature evaluation. The colored bars represent the mean scores and the error bars indicate the standard deviations.

6.2 Overall Feedback

Since we recruited experts from a diverse array of domains, we were interested in understanding which features were most and least useful for each domain.

The experts universally expressed interest in and found much value in the indicator the score features, albeit with concerns regarding accuracy. *ES: Great idea, as a coach I would pick the most relevant positions the [user] needs to match step by step. ED: It would be useful since it can help me in real time on if I'm performing the technique poorly. So for example, if my legs are spread too wide, the video and avatar will not progress until I perform the action correctly. It will force me to narrow them before we move to the next step. EG: The most useful feature for golf. I like that it incorporates all the different elements of your swing. How your knees bend and move during setup and the swing. I like that it would tell me in real time if my shoulders, wrists and other parts are correctly set up for the swing in real time. I would add more joints, like hips if you can. EY: I really like it, it's fun. If you can maybe select where you want it to be evaluating you, like only evaluating my arms, or legs, or feet, etc would be very great. EM: Would use the feature for sure. To perform a demo of a movement then ask someone to mimic the movement would be super useful in this way. EF: I like being able to synchronize with the instructor's avatar. Since it follows joints, it will be very useful for fitness.*

As for body-based navigation, all experts again expressed interest in the feature but found it largely geared towards beginners, as intermediate and advanced learners would likely need to practice more fluidly. *ED: this is probably better for beginners to learn each step, because more experienced dancers would want to perform more fluidly. ES: Could be difficult to use as you level up since the techniques should be done as one movement, not broken up into steps. EG: This would allow you to see how your body is moving compared to the avatar in a slow motion fashion, which is probably a better way to learn a golf swing if you haven't learned yet. Most people have trouble with golf by overswinging, they hit the ball as hard as they can. If you go step by step and in slow motion, your body will learn the motions and positions more accurately, i.e. not smashing the ball, and you'll learn through muscle memory. EF: Being able to show students that they're doing the proper technique step by step is very useful and especially through different positions. EM: I don't want to see how students are positioned at the start and end of exercise, I want to see and teach them step by step where they should be located and angled through the whole exercise which this helps a lot for.*

6.3 Domain Specific Feedback

6.3.1 ***Yoga***. While greatly interested in the indicator, score, and body-based navigation features, the yoga expert found little value for trajectory and footprint since yoga requires little movement. *EY: "I liked that [footprints] were easy to synchronize. I could just look down and plant my feet the same as the avatar. But we don't move much in Yoga so I'd only do it once (...) since we hold the poses for an amount of time, it's not really important to see the trajectory of the body. Often there isn't even a trajectory to see. I would rather just see the outline of the pose".* However, they explain situations in which the head-gaze feature would be useful. *EY: sometimes seeing*

the head-gaze would be useful for poses where it's important to keep your head still or facing a certain direction. For example the Warrior 2 Pose in Yoga needs you to face forwards while your body twists, which for beginners is difficult to do. But seeing the avatar's head-gaze not moving would make it easier to understand just how little the head should move.

6.3.2 ***Dancing.*** As we expected, the dancing expert found the footprints and trajectory features the most useful. 1st person footprints more than 3rd person, 3rd and 1st person trajectory, though 1st person trajectory had it's value diminished due to the limited Hololens FOV. *ED: I can't really see it well [due to Hololens FOV]. Footprints are probably the most useful feature in first person, especially for footwork-heavy dances. Third person is much better for individual learning and group learning.* The expert also appreciated the head-gaze feature, albeit in combination with trajectory. *ED: "During group dancing, we should all be synchronized. Viewing the trajectory and head-gaze movements together could help us all while practicing to make sure we stay synchronized, which is critical to certain kinds of dances".*

6.3.3 ***Soccer.*** The soccer expert found the trajectory very useful for visualizing shooting the ball. *ES: "It all aligns for how it should look like. You want the trajectory, especially for soccer, of the body the head and the foot and where the ball ends up. If i was teaching, i would want to draw a line from my foot starting to swing, to hitting the ball to where the ball goes. I would tell a kid to draw a line [with their dominant leg] from where the ball is to where they wanted it to go. I'd want to be able to select which limbs to track rather than the whole body at once and i want to select when to start drawing the trajectory and when to not, like when I'm explaining something".* They also enjoyed head-gaze for understanding and visualizing different methods of heading a ball. *ES: "[Head-gaze is] useful for visualizing headers, since there's different kinds of headers (ex: straight, glance, etc). You could visualize the head rotation and position for each of those as well as how to sync to them [with 1st person]".*

6.3.4 ***Golf.*** The golf expert found our system best for learning how to set up a golf swing, both by using the first person footprints for seeing where to stand *EG: "[Footprints] are useful during setup to put your feet in the right positions"* and head-gaze to remind users to stare at the ball during the entire swing. *EG: "The key thing in golf is to have your head looking at the ball throughout the whole swing which [head-gaze] certainly does".* The expert also liked the third person trajectory for visualizing the overall circular swing. *EG: "I liked how it shows the natural arc of a golf swing and shows what it should look like as an example. I wish it would consider wrist positions, body positions, for the swing as well".*

6.3.5 ***Fitness.*** The fitness expert enjoyed first and third person POVs of footprints and trajectory to prepare for and performing exercises. *EF: "Seeing the path of the hands, for movements and exercises like dead-lifts would be great. Could show the movement of the exercise [and therefore the equipment] without [the avatar] actually showing the equipment". EF: "It's useful to see the path of footprints. For example for walking lunges, zumba, etc."*

6.3.6 ***Martial Arts.*** Similar to the fitness expert, the martial arts expert enjoyed first and third person POVs of trajectory and footprints the most. *EM: "I really liked [trajectory], it's the most worthwhile and [the feature I'm most] likely to use. It's best for watching someone do an exercise and even better [combined with footprints] to see the path of the technique".* Unlike the fitness expert they also found head-gaze useful context for learners. *EM: "it would be great to offer the student the ability to see where I"m looking and match head motions".*

6.4 Expert Insights

6.4.1 ***Embodiment of Professionals.*** The golfing, dancing, and soccer experts each also expressed an interest in moving beyond YouTube tutorials and instead embodying professionals and celebrities. The soccer expert expressed excitement at the ability to embody professional soccer players *S1: "One of the biggest things for soccer is you want to be the professional, you want to know what the professional sees and what they do based on the information at hand. The best way to relay that to someone is to let them be the athlete. It would also be very insightful to see how quickly they dribble the ball and their footwork especially for learning tricks".* The dancing expert echoed this sentiment *ED: "If I had videos of Beyoncé on stage I'd love to learn her dances then perform it like she does on stage. I'm really interested in seeing her 1st person perspective".* Finally, the golfing expert also communicated interest in the embodiment of professional golfers *EG: "The swing might be hard to see in first person but it'd be useful to see how Tiger Woods sets up [to swing]".*

6.4.2 ***More Scalable Education.*** Although our system was designed for 1-1 education and can only accommodate one user at a time, our experts universally expressed great interest and excitement for using our system during group lessons. Most experts explain that, financial concerns aside, they would equip a group of students with a Hololens each and have everyone watch the 3rd person avatar's movements as a group. *ES: "I would use 3rd person as a group example for the kids (...) In general, third person is useful for group explanations, while first person is better for practicing techniques and individual training.".* They imagine that students would walk around and observe the movements from various angles in 3D while the expert explains what's happening or what to pay attention to. Experts would be in control of which features are active at what time (footprint, trajectory, etc) and would especially use them as additional visual context for their oral explanation. With the demonstration complete, they would then break the students off to practice individually. They believe that each student would be able to practice, learn and improve more independently due to our system's ability to 1) enable students to replay the 3D lesson (eg: for details they may have missed), 2) to embody the instructor's avatar (for a first-person perspective of the lesson) and 3) most importantly of all, constantly outputting real-time feedback tailored for each student individually (which helps them improve without the need for intervention or feedback from the expert). *ES: "you could practice that mechanic over and over again building the muscle memory and body motions. If accurate, this could enable automatic training for soccer techniques because it gives you real-time feedback".* Furthermore, they expressed interest in monitoring student scores as a whole, for an overall picture of class performance and to tailor

interventions. *ED: "If it could give me feedback on the scores of how the kids are doing, that would be amazing because it would let me check which techniques each kid is having difficulty with or group kids with similar difficulties to teach together in smaller groups"*. Finally, some experts also mentioned using our system as a method for supplemental or remote learning. *"ES: "it would be great to give to kids as practice/homework to do while away from the instructor or learning remotely". EF: "Would be great to use for teaching people how to use their own home gyms for fitness training". EM: If we were practicing a technique and i was teaching remotely, it would be great to offer the student the ability to see where I"m looking and match head motions.*

6.4.3 ***Comparison to Traditional Instruction***. Our golfing expert pointed out how indoor golfing simulations have become ubiquitous, capable of providing accurate, useful and immediate feedback, enabling golfers to practice their craft more effectively, though the feedback is limited to the **outcome** of a golfer's actions and not the action itself. *EG: "The tech for indoor driving ranges are really good, they can tell you the angle of the ball, speed, wind direction, spin on the ball, etc. They're very accurate, but they don't tell you anything about your actual swing, your body positions or your motions"*. They then explain how our system can combine with indoor golfing simulations to provide a more holistic and complete evaluation of a user's golfing performance. *G1: "Your system on the other hand helps tremendously for learning how to setup for a golf swing and gives you feedback on the swing itself, which is something currently missing in the market. We have accurate info for what happens to the ball after you swing, but nothing for what happens prior. Combining them is the best of both worlds"*.

6.5 Suggestions for Feature Expansion

Our experts suggested a few different improvements for our features but for the most part they requested additional contextual cues. *ES: "Would also be nice to have different color trajectories for each stage of the technique like before the leg hits the ball, after leg hits the ball, etc". ED: "An arrow or something to show the direction in which the trajectory is moving would be super helpful in quickly identifying the actual trajectory of the limb if the trajectory gets messy or I pause it". ED: "being able to see the whole trajectory of the whole movement at once instead of the past few second would be useful to prepare the rest of our body for the upcoming motions based on the trajectory that we can see the limbs needs to move in". EY: "Could you show the degree of the angle of the joints on the 3D avatar and my own [Kinect] avatar? Like if I lean to the side but keep my legs still, can you show how many degrees I've leaned or angled? I'd like that info" EF: "Would be useful to be able to click the trajectory and see where they were at that time, especially for viewing what their other limbs were doing in that moment and even click to rewind and fast forward". ED: "Numbering the footprints or even highlighting the them with the tempo/lyrics, etc so I can visualize the order in which i need to step much better"*. The dancing instructor also requested the ability to modify the tempo of the input music and have the 3D avatar's animation synchronize automatically. *ED: "Sometimes we synchronize [our dances] with either the lyrics, basses or instruments. I'd like it if we could visualise those alongside the avatar. Like showing which instruments are playing at what part of the dance or the music tempo"*. Expanding on synchronized dancing, the expert requests the ability modify the music and have the avatar's dance change accordingly.*ED: "It would be useful if I could change the tempo of the music from the video and see the avatar dancing the same moves but automatically adjusted to the different tempo I set. Like I could quickly see a preview of the dance to the new tempo through the avatar"*.

6.6 Technical Limitations

While each domain expert had varying opinions on our system overall, as well as each of the features, some opinions were expressed by all experts. Most commonly, experts found the limited FOV of the Hololens 2 challenging to deal with, especially for the first-person features. For example, while our martial arts expert enjoyed the trajectory feature to *EM: "visualize a punching or kicking"* they expressed concern and frustration with following more than one movement at a time using the first person perspective. *EM: "I can only sync one hand at a time since I can't really view them both at once"*. Experts from domains that required the use of equipment, like golf and soccer, expressed disappointment from the lack of equipped equipment by the 3D avatar. *ES: "I wish be you could use a real ball in with the headset to give that physical feedback to the students"*. On the other hand, the fitness expert expressed less disappointment at the lack of equipment, claiming that trajectory is still useful to see how the exercise is performed, even without the avatar equipping equipment. *EF: "Seeing the path of the hands, for exercises like dead-lifts would be great. This would show the movement of the exercise and the equipment"*. Finally, all the participants expressed concern over inaccuracy of Kinect body tracking for the indicator, score and body-based navigation features causing difficulty to synchronize their body to the 3D avatar.

7 Future Work

Object Detection. In our study, some participants mentioned that it was difficult to follow the avatar instructions for videos that use equipment such as tennis and soccer. To address this, we could attach 3D object models of equipment to the avatar. For example, we can attach a virtual tennis racket to the avatar's hand like *RealityCanvas* [57] and rotate the racket based on the hand rotation. Also, we could detect the objects in the video using object detection techniques. For example, if we could detect the 3D position of the soccer ball, users can watch where the instructor kicks the ball and how the ball moves.

Also, while we learn sports that use instruments, we often use those instruments as well, such as tennis rackets and soccer balls. If we can track the real object that the user is using, we can show the user a more immersive instructional experience. For example, for the soccer instruction scenario, by synchronizing the avatar's animation with the position of a soccer ball, user can see the exact point where an instructor kicks the ball.

Verbal Information. Verbal information is important to understand the detailed information. Inspired by previous works, such as *Reality Talk* [30], we envision enhancing the avatar representation based on what the instructor is talking about. We can leverage speech recognition techniques such as *Google Cloud Speech-to-Text*. By using the verbal information, we could highlight the specific body position or show additional information about the avatar

based on what they are talking about. For example, when the instructor is talking about specific body parts, we can highlight those parts. When the instructor explains domain-specific words, we can display some related figures and text to explain them.

2D Visualized Information. Some 2D videos incorporate on-screen visualizations to enhance clarity and encouragement. For example, they use time and numbers to encourage the user and use highlighting and arrows to emphasize body parts. We could extract that information from the video by using optical character recognition (OCR) techniques, such as *Google Cloud Vision API* and convert it into 3D could help the user understand the instructions and be encouraged. Also, some videos refer to other videos, such as professional videos or their own previous videos. Combining extracting body motion from several videos and showing several avatars could display more detailed information.

Avatar Representation. Previous studies, such as Tsuchida et al. [50], have explored how avatar appearance influences instructional effectiveness. However, our research shifts focus to the movements, spatial positioning, and temporal dynamics of avatars rather than their physical appearance. In this study, we employed a 3D avatar sourced from Mixamo, which did not visually resemble the instructor featured in the video. While this approach was sufficient for our purposes, future work could leverage 3D reconstruction technologies to create avatars with textures that closely mirror the instructor's likeness. This level of fidelity could heighten users' sense of presence and engagement by making them feel as though the instructor is physically present.

Ethical Considerations. While the Video2MR system ensures that all 3D mesh data and feedback data are processed in real-time without storage or sharing, we recognize that ethical concerns related to user privacy and security must remain central to the deployment of such technologies. As the system tracks and analyzes user body movements, it is critical to ensure that all data processing occurs locally and solely for the purpose of immediate interaction. This safeguards against risks of data breaches or misuse. Moreover, there is potential for biases in motion capture systems, particularly for individuals with diverse body types, physical abilities, or movement styles. Such biases could result in inequitable or inaccurate feedback, raising issues of fairness and inclusivity. Future iterations of the system should address these concerns by incorporating models trained and tested on diverse datasets, alongside adding mechanisms in the tool for user feedback to refine system accuracy and equity.

Diverse Participant Pool. While our study demonstrated the effectiveness of the system for users familiar with video-based instruction, as indicated by the participants' high average familiarity rating of 6.42/7, we recognize the need to explore its impact on a broader range of users. Specifically, evaluating the system's usability and engagement potential for individuals with limited prior experience in video-based learning presents a promising direction for future research.

8 Conclusion

In this paper, we developed Video2MR, which enhances and automatically generates mixed reality instructions utilizing extracted human motion from 2D instructional videos. Video2MR has four design elements: 1) Comparison, 2) Visualization, 3) Navigation, 4) Reposition, and we implemented several features based on them. We conducted two evaluations to evaluate the usefulness of Video2MR compared with just using 2D videos. We conducted a user study with 12 participants and confirmed that our system can enhance co-presence, engagement, and fun. Also, we found using video in the background of the avatar could help users follow the instructions. Through the expert reviews with six participants, we confirmed that our implemented features are useful for a variety of physical exercise domains, gathered unique qualitative insights like how experts wish to embody celebrities and professionals, learned that there is a desire to enhance current instructional methods with the concepts and features presented by our system and finally that Video2MR enables more scalable instruction by empowering beginners and novices to learn more independently. Additionally, we mentioned how we can enhance the experience through object detection, verbal detection, and improving avatar representation as a future work.

Acknowledgments

This research was funded part by the NSERC Discovery Grant RGPIN-2021-02857 and JST PRESTO Grant Number JPMJPR23I5. We also thank all of the participants for our user study.

References

[1] Jindřich Adolf, Peter Kán, Benjamin Outram, Hannes Kaufmann, Jaromír Doležal, and Lenka Lhotská. 2019. Juggling in vr: Advantages of immersive virtual reality in juggling learning. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*. 1–5.

[2] Judith Amores, Xavier Benavides, and Pattie Maes. 2015. ShowMe: A Remote Collaboration System That Supports Immersive Gestural Communication. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1343–1348.

[3] 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.

[4] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[5] Marion Buchenau and Jane Fulton Suri. 2000. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. 424–433.

[6] Subramanian Chidambaram, Sai Swarup Reddy, Matthew Rumple, Ananya Ipsita, Ana Villanueva, Thomas Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2022. EditAR: A Digital Twin Authoring Environment for Creation of AR/VR and Video Instructions from a Single Demonstration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 326–335.

[7] Christopher Clarke, Doga Cavdir, Patrick Chiu, Laurent Denoue, and Don Kimber. 2020. Reactive video: adaptive video playback based on user motion for supporting physical activity. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 196–208.

[8] Caleb Conner and Gene Michael Poor. 2016. Correcting exercise form using body tracking. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3028–3034.

[9] DeepMotion. 2023. DeepMotion. https://www.deepmotion.com/

[10] Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N Balasubramanian, Bharathi Callepalli, and Ayon Sharma. 2022. Pose tutor: an explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3540–3549.

[11] Daniel Eckhoff, Christian Sandor, Christian Lins, Ulrich Eck, Denis Kalkofen, and Andreas Hein. 2018. TutAR: augmented reality tutorials for hands-only procedures. In *Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. 1–3.

[12] 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.

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*. 2334–2343.

[14] Mehrad Faridan, Bheesha Kumari, and Ryo Suzuki. 2023. ChameleonControl: Teleoperating Real Human Surrogates through Mixed Reality Gestural Guidance for Remote Hands-on Classrooms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 13 pages.

[15] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. 2021. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9919–9928.

[16] Natsuki Hamanishi, Takashi Miyaki, and Jun Rekimoto. 2019. Assisting viewpoint to understand own posture as an avatar in-situation. In *Proceedings of the 5th International ACM In-Cooperation HCI and UX Conference*. 1–8.

[17] Natsuki Hamanishi and Jun Rekimoto. 2020. Poseasquery: Full-body interface for repeated observation of a person in a video with ambiguous pose indexes and performed poses. In *Proceedings of the Augmented Humans International Conference*. 1–11.

[18] Natsuki Hamanishi and Jun Rekimoto. 2021. Motion-specific browsing method by mapping to a circle for personal video Observation with Head-Mounted Displays. In *Proceedings of the Augmented Humans International Conference 2021*. 240–250.

[19] Ping-Hsuan Han, Kuan-Wen Chen, Chen-Hsin Hsieh, Yu-Jie Huang, and Yi-Ping Hung. 2016. Ar-arm: Augmented visualization for guiding arm movement in the first-person perspective. In *Proceedings of the 7th Augmented Human International Conference 2016*. 1–4.

[20] Ping-Hsuan Han, Yang-Sheng Chen, Yilun Zhong, Han-Lei Wang, and Yi-Ping Hung. 2017. My Tai-Chi coaches: an augmented-learning tool for practicing Tai-Chi Chuan. In *Proceedings of the 8th Augmented Human International Conference*. 1–4.

[21] Thuong N Hoang, Martin Reinoso, Frank Vetere, and Egemen Tanin. 2016. Onebody: remote posture guidance system using first person view in virtual environment. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. 1–10.

[22] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J Quinn. 2021. Adaptutar: An adaptive tutoring system for machine tasks in augmented reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[23] Elena Iannucci, Zhutian Chen, Iro Armeni, Marc Pollefeys, Hanspeter Pfister, and Johanna Beyer. 2023. ARrow: A Real-Time AR Rowing Coach. (2023).

[24] Keiichi Ihara, Mehrad Faridan, Ayumi Ichikawa, Ikkaku Kawaguchi, and Ryo Suzuki. 2023. HoloBots: Augmenting Holographic Telepresence with Mobile Robots for Tangible Remote Collaboration in Mixed Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[25] Atsuki Ikeda, Dong-Hyun Hwang, Hideki Koike, Gerd Bruder, Shunsuke Yoshimoto, and Sue Cobb. 2018. AR based Self-sports Learning System using Decayed Dynamic TimeWarping Algorithm.. In *ICAT-EGVE*. 171–174.

[26] Yao-Fu Jan, Kuan-Wei Tseng, Peng-Yuan Kao, and Yi-Ping Hung. 2021. Augmented Tai-Chi Chuan Practice Tool with Pose Evaluation. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 35–41.

[27] 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.

[28] Nicholas Katzakis, Jonathan Tong, Oscar Ariza, Lihan Chen, Gudrun Klinker, Brigitte Röder, and Frank Steinicke. 2017. Stylo and handifact: Modulating haptic perception through visualizations for posture training in augmented reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*. 58–67.

[29] Felix Kosmalla, Fabian Hupperich, Anke Hirsch, Florian Daiber, and Antonio Krüger. 2021. VirtualLadder: Using Interactive Projections for Agility Ladder Training. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[30] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[31] 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.

[32] Jingyuan Liu, Nazmus Saquib, Zhutian Chen, Rubaiat Habib Kazi, Li-Yi Wei, Hongbo Fu, and Chiew-Lan Tai. 2022. PoseCoach: A Customizable Analysis and Visualization System for Video-based Running Coaching. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[33] Ruofan Liu, Erwin Wu, Chen-Chieh Liao, Hayato Nishioka, Shinichi Furuya, and Hideki Koike. 2023. PianoSyncAR: Enhancing Piano Learning through Visualizing Synchronized Hand Pose Discrepancies in Augmented Reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 859–868.

[34] Zoe Marquardt, João Beira, Natalia Em, Isabel Paiva, and Sebastian Kox. 2012. Super Mirror: a kinect interface for ballet dancers. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 1619–1624.

[35] Takashi Matsumoto, Erwin Wu, and Hideki Koike. 2022. Skiing, Fast and Slow: Evaluation of Time Distortion for VR Ski Training. In *Proceedings of the Augmented Humans International Conference 2022*. 142–151.

[36] Fariba Mostajeran, Frank Steinicke, Oscar Javier Ariza Nunez, Dimitrios Gatsios, and Dimitrios Fotiadis. 2020. Augmented reality for older adults: exploring acceptability of virtual coaches for home-based balance training in an aging population. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[37] Thammathip Piumsomboon, Gun A Lee, Jonathon D Hart, Barrett Ens, Robert W Lindeman, Bruce H Thomas, and Mark Billinghurst. 2018. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[38] Thammathip Piumsomboon, Gun A Lee, Andrew Irlitti, Barrett Ens, Bruce H Thomas, and Mark Billinghurst. 2019. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–17.

[39] Zelai Saenz-de Urturi and Begonya Garcia-Zapirain Soto. 2016. Kinect-based virtual game for the elderly that detects incorrect body postures in real time. *Sensors* 16, 5 (2016), 704.

[40] Christian Schönauer, Kenichiro Fukushi, Alex Olwal, Hannes Kaufmann, and Ramesh Raskar. 2012. Multimodal motion guidance: techniques for adaptive and dynamic feedback. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 133–140.

[41] Yoones A Sekhavat and Mohammad S Namani. 2018. Projection-based AR: Effective visual feedback in gait rehabilitation. *IEEE Transactions on Human-Machine Systems* 48, 6 (2018), 626–636.

[42] Alessandra Semeraro and Laia Turmo Vidal. 2022. Visualizing instructions for physical training: Exploring visual cues to support movement learning from instructional videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[43] Lucchas Ribeiro Skreinig, Ana Stanescu, Shohei Mori, Frank Heyen, Peter Mohr, Michael Sedlmair, Dieter Schmalstieg, and Denis Kalkofen. 2022. AR Hero: Generating interactive augmented reality guitar tutorials. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 395–401.

[44] Rajinder Sodhi, Hrvoje Benko, and Andrew Wilson. 2012. LightGuide: projected visualizations for hand movement guidance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 179–188.

[45] Ryo Suzuki, Rubaiat Habib Kazi, Li-Yi Wei, Stephen DiVerdi, Wilmot Li, and Daniel Leithinger. 2020. Realitysketch: Embedding responsive graphics and visualizations in AR through dynamic sketching. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 166–181.

[46] Yuto Suzuki, Daisuke Sakamoto, and Tetsuo Ono. 2023. Gino. Aiki: Mixed Reality-based Physical Motor Skill Training in Aikido. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 519–524.

[47] Santawat Thanyadit, Parinya Punpongsanon, and Ting-Chuen Pong. 2019. ObserVAR: Visualization system for observing virtual reality users using augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 258–268.

[48] Atima Tharatipyakul, Kenny TW Choo, and Simon T Perrault. 2020. Pose estimation for facilitating movement learning from online videos. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–5.

[49] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating remote instruction of physical tasks using bi-directional mixed-reality telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 161–174.

[50] Shuhei Tsuchida, Haomin Mao, Hideaki Okamoto, Yuma Suzuki, Rintaro Kanada, Takayuki Hori, Tsutomu Terada, and Masahiko Tsukamoto. 2022. Dance practice system that shows what you would look like if you could master the dance. In *Proceedings of the 8th international conference on movement and computing*. 1–8.

[51] Laia Turmo Vidal, Elena Márquez Segura, Christopher Boyer, and Annika Waern. 2019. Enlightened yoga: Designing an augmented class with wearable lights to support instruction. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 1017–1031.

[52] Laia Turmo Vidal, Hui Zhu, and Abraham Riego-Delgado. 2020. Bodylights: Open-ended augmented feedback to support training towards a correct exercise execution. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[53] Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2013. Motionma: motion modelling and analysis by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1309–1318.

[54] Cheng Yao Wang, Qian Zhou, George Fitzmaurice, and Fraser Anderson. 2022. VideoPoseVR: Authoring Virtual Reality Character Animations with Online Videos. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS (2022), 448–467.

[55] 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.

[56] Chengshuo Xia, Xinrui Fang, Riku Arakawa, and Yuta Sugiura. 2022. VoLearn: A Cross-Modal Operable Motion-Learning System Combined with Virtual Avatar and Auditory Feedback. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2022), 26 pages.

[57] Zhijie Xia, Kyzyl Monteiro, Kevin Van, and Ryo Suzuki. 2023. RealityCanvas: Augmented Reality Sketching for Embedded and Responsive Scribble Animation Effects. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[58] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977* (2018).

[59] Shuo Yan, Gangyi Ding, Zheng Guan, Ningxiao Sun, Hongsong Li, and Longfei Zhang. 2015. Outsideme: Augmenting dancer's external self-image by using a mixed reality system. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 965–970.

[60] Shuainan Ye, Zhutian Chen, Xiangtong Chu, Yifan Wang, Siwei Fu, Lejun Shen, Kun Zhou, and Yingcai Wu. 2020. Shuttlespace: Exploring and analyzing movement trajectory in immersive visualization. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 860–869.

[61] Xingyao Yu, Katrin Angerbauer, Peter Mohr, Denis Kalkofen, and Michael Sedlmair. 2020. Perspective matters: Design implications for motion guidance in mixed reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 577–587.

[62] Ziyi Zhao, Sena Kiciroglu, Hugues Vinzant, Yuan Cheng, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. 2022. 3d pose based feedback for physical exercises. In *Proceedings of the Asian Conference on Computer Vision*. 1316–1332.

[63] Zhongyi Zhou, Anran Xu, and Koji Yatani. 2021. Syncup: Vision-based practice support for synchronized dancing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–25.