Virtual Reality

https://doi.org/10.1007/s10055-026-01395-2

Article in Press

Xarnold: assessing the role of virtual humans in immersive fitness environments with aligned flat, multi-view, and 3D parametric visual instruction guidance

Received: 28 November 2025

Accepted: 5 May 2026

Published online: 24 June 2026

Cite this article as: Sirocchi C., Stacchio L., Migliorelli L. *et al.* Xarnold: assessing the role of virtual humans in immersive fitness environments with aligned flat, multi-view, and 3D parametric visual instruction guidance. *Virtual Reality* (2026). https://doi.org/10.1007/s10055-026-01395-2

Claudio Sirocchi, Lorenzo Stacchio, Lucia Migliorelli, Emanuele Frontoni & Adriano Mancini

Xarnold: assessing the role of virtual humans in immersive fitness environments with aligned flat, multi-view, and 3D parametric visual instruction guidance

XaRNold: Assessing the Role of Virtual Humans in Immersive Fitness Environments with aligned Flat, Multi-View, and 3D Parametric Visual Instruction Guidance

Claudio Sirocchi[1 [0009−0009−8258−2158]*], Lorenzo Stacchio[3 [0000−0002−9341−7651]*], Lucia Migliorelli[2 [0000−0003−2388−1501]], Emanuele Frontoni[3 [0000−0002−8893−9244]], and Adriano Mancini[1 [0000−0001−5281−9200]]

[1] Università Politecnica delle Marche, Department of Information Engineering, Ancona, Italy
[2] Università degli Studi di Teramo, Department of Political Science, Teramo, Italy
[3] University of Macerata, Department of Political Science, Communication, and International Relations, Macerata, Italy
c.sirocchi@pm.univpm.it, lorenzo.stacchio@unimc.it, lmigliorelli@unite.it, emanuele.frontoni@unimc.it, a.mancini@staff.univpm.it

Abstract. Recent advances in Extended Reality (XR) allow the creation of immersive fitness and virtual coaching systems. However, it remains unclear whether immersive visualization modalities influence users' perception, understanding, and confidence in performing exercises. Moreover, it is unclear whether this improves those with respect to a classical 2D visualization. In this work, we investigate the role of a parametric 3D human model in enhancing workout perception and social presence within immersive fitness environments. To this date, we have designed a modular XR system, named *XaRNold*, that allows the visualization of single/multi-view video, and a parametric 3D model called Skinned Multi-Person Linear Model (SMPL) showing how to perform a physical exercise. We conducted a controlled user study (N=30) to compare participants' cognitive load/understanding, technology acceptance, usability, and social perceptions/engagement across visualization modes. We employ the Fit3D dataset, which provides aligned 2D, multi-view, and related 3D poses data for diverse physical exercises. The results indicate that a parametric virtual human model provides a favorable trade-off between the measured constructs, offering a clearer understanding of why and how avatar-based guidance can be used and paving the way for more effective and empathic virtual fitness systems.

Keywords: Immersive Workout · Visual Guidance · Social Virtual Reality · Virtual Humans · Extended Reality

* Claudio Sirocchi and Lorenzo Stacchio should be considered as joint first authors.

1 Introduction

Over the past decade, the emergence of the Metaverse and Extended Reality (XR) technologies has redefined how humans interact with digital environments across various sectors. XR has enabled novel forms of immersion, embodiment, and real-time interaction, reshaping domains such as education [47], healthcare [34], and entertainment [23]. These immersive environments offer spatial presence, contextual reactivity, and the ability to simulate or augment real-world scenarios, making them particularly attractive for applications where visualization and physical interaction are crucial.

Among these domains, physical activity has emerged as a promising XR use case: compared to traditional screen-based solutions, immersive systems can provide embodied interaction, spatial awareness, and adaptive feedback, supporting more interactive and potentially more engaging exercise experiences [2, 43, 21, 28]. Indeed, prior studies have shown that XR fitness can support training, rehabilitation, and performance-oriented activities in immersive settings, often reporting benefits in motivation, engagement, and adherence. These effects were even greater when paired with an embodied virtual instructor [43, 28, 21, 39].
However, the majority of this literature has mainly emphasized affective and experiential benefits, including enjoyment, immersion, reduced perceived exertion, and sustained interest in exercising [41, 14, 28, 21, 39]. Therefore, there is less exploration of how different visual guidance modalities influence users' understanding of physical movements. This also includes virtual instructor demonstrations, which may also have an impact on the social dimension [19], shaping how the user feels supported and accompanied during exercise.
This leaves an open question: *are gains in understanding and performance primarily a byproduct of gamification and motivation, or does the immersive and socially embodied nature of visualization play a more direct perceptual and motor role?* Isolating these dimensions (visual guidance and social presence) provides an opportunity to use XR fitness as a testbed to identify which factors most effectively support exercise comprehension relative to a classical flat video. In this perspective, comparing widely adopted exercise visualizations (i.e., flat video, multi-view representations, and embodied 3D guidance) under aligned exercise content can help clarify which factors most effectively support exercise comprehension in immersive settings.
To the best of our knowledge, only a limited number of prior works have moved in this direction [9, 13, 10]. However, these studies remain focused on specific scenarios and do not yet provide a unified comparison framework based on aligned stimuli, scalable exercise support, and intermediate visualization conditions. This outlines the need for a modular and controlled XR setting able to disentangle how different guidance modalities influence exercise comprehension and perceived support.

To address these limitations, we move beyond isolated factor evaluations and provide an ecologically grounded assessment of how immersive visualization and virtual social presence jointly influence exercise comprehension and perceived support. Specifically, we explore whether parametric 3D human representations

(when compared to aligned flat 2D or multi-view video guidance) can improve users' understanding of physical movements and enhance their sense of companionship during training in VR. To this end, we developed *XaRNold* [4] (eXtended Reality for teAching and social iNteraction in Learning physical exercises), a modular immersive XR fitness system that supports multiple synchronized visualization modalities for exercise instruction, also including an automatic workout scheduler. These include traditional 2D single-view video, a composite multi-view layout, and a fully interactive 3D avatar rendered using the Skinned Multi-Person Linear Model (SMPL) parametric human model. All these visualizations were built on top of the Fit3D dataset [17], which provides temporally aligned exercise poses, allowing us for a direct comparison of user responses across conditions.

We employed *XaRNold* to define an experimental setting, to answer the following Research Questions (RQs):

- **RQ1**: When exercise content and showing timing are held constant, how do the immersive visualization modalities compare to classical flat 2D, in terms of participants' general exercise learning?
- **RQ2**: How do the three visualization modalities differ in terms of perceived instructor social presence, comfort, embodiment, and affective support during XR-based exercise instruction?
- **RQ3**: Does a parametric 3D avatar improve perceived clarity, workload, and social presence compared to 2D video modalities?

We employed this system in a controlled within-subjects user study, where (N=30) participants performed guided exercises (i.e., a simple workout) under each visualization mode. By isolating the effects of visual modality in a realistic workout scenario, we seek to provide evidence on whether immersive 3D trainers offer effective advantages beyond engagement or novelty. It is worth noticing that, in this study, we did not treat motor learning as an outcome (e.g., execution correctness or retention). We instead answered our RQs, investigating:

- (RQ1, RQ3) Learning and cognitive self-perceived outcomes related to the comprehension of the workout exercises (i.e., whether participants felt they understood how to perform the movement);
- (RQ2) User experience factors such as usability, task-load index, and technology acceptance;
- (RQ2, RQ3) Affective and social perceptions, including perceived presence and companionship;

The obtained results indicate that the 3D avatar visual stimuli provide a favorable trade-off between instructional clarity, perceived usefulness, and social presence, while classical 2D single-video remains the most familiar baseline, and multi-view videos tend to sit in-between but can introduce split-attention costs. This evidence fills a specific literature gap: unlike many XR fitness studies that

[4] An exemplar video demo of *XaRNold* and the experimental settings are available in the supplementary files.

analyzed engagement with heterogeneous stimuli, our aligned-stimulus comparison isolates visual guidance modality as the primary factor, offering a clearer understanding of why and how avatar-based guidance can be used, instead of resorting to classical 2D baselines.

In summary, the main contributions of this work can be summarized as follows:

- A modular XR fitness framework (XaRNold) and its implementation, which enables controlled cross-modality comparison delivering the same temporally aligned instructional stimulus across three aligned modalities (to isolate visualization factors while keeping exercise content constant).
- A unified stimulus pipeline based on shared motion sources, describing how a single exercise representation can be transformed into the three delivery formats, ensuring consistent timing and content to support fair within-subject comparisons.
- A mixed-methods within-subject evaluation protocol for XR exercise guidance, reporting a controlled user study (N=30) combining validated quantitative measures with qualitative feedback to explain observed differences and derive design implications.
- Empirical evidence and design implications showing how different visualization modalities trade off instructional clarity, perceived usefulness, cognitive load, usability, and social presence in XR fitness guidance.

This manuscript is organized as follows. In Section 2, we review related literature on immersive fitness, visual guidance, and virtual embodiment. Section 3 details the architecture and features of the *XaRNold* system. Section 4 describes the experimental setup, participants, and apparatus. Results and statistical analyses are presented in Section 5. Finally, Section 6 discusses the implications of the obtained results, this work's limitations, and outlines directions for future research.

2 Related Work

XR and, in particular, VR technologies are increasingly adopted in sports science and rehabilitation to enhance engagement, motivation, and exercise comprehension [30, 32]. XR emerged as a particularly promising application because embodied interfaces, spatial awareness, and adaptive feedback can provide more interactive experiences than traditional screen-based platforms [2, 60, 43, 21, 28]. Nowadays, existing systems supports a wide range of activities, including training in virtual gyms, multiplayer fitness experiences, rehabilitation, and performance optimization, without being constrained by physical location [43, 28]. Recent research has indeed shown a growing number of XR fitness systems exploiting sensory environments and gamified designs [2, 37, 21, 3], with most contributions developed in VR due to its ability to provide immersion, engagement, and personalized feedback in real time [28, 21, 58]. These studies often report increased

interest in exercising and stronger adherence-related effects with respect to traditional approaches [28, 21, 39]. Moreover, including social and co-presence effects in VR exercise settings, showing that the presence of virtual others (e.g., avatars or companions) can enhance performance and enjoyment [41]. This was also observed in studies focusing on embodiment manipulations (e.g., muscular avatars or virtual runners) to influence user perception and behavior, measuring constructs related to embodiment, presence, and motivation, highlighting how virtual social representation can directly influence immersive experience [41, 14, 54, 27, 19].

Despite these relevant findings, few studies systematically compare different visualization modalities with respect to the dimensions that are relevant for guided exercise systems, including instructional and perceptual clarity, perceived social presence of the virtual instructor, and usability, workload, and acceptance factors. As an example, [60] introduced a VR training system integrating a headset, sensors, and a virtual coach to help beginners learn correct exercise postures, reporting high usability, motivation, and perceived learning effectiveness. On the same line, [58] proposed an AI-based adaptive VR sports system that uses deep reinforcement learning to train virtual coaches for personalized and empathy-oriented sports guidance. Complementary findings also indicate that employing a muscular embodied avatar can enhance self-perception and execution performance [14]. Such works suggest that visual demonstrations can make kinematics, joint alignment, posture, and movement patterns easier to interpret, especially when movements are inspected from informative perspectives [52, 50, 59]. At the same time, these studies mainly focus on engagement, motivation, and general exercise learning and do not compare how different forms of visual guidance affect exercise comprehension. Moreover, they often under-examine the social dimension (how the co-presence of a 3D virtual instructor may shape the perceived clarity, support, and companionship of immersive exercise guidance) [57, 14]. This recent literature outlines three outcome families that are especially relevant for XR-based exercise guidance: (i) exercise comprehension and perceptual clarity of the demonstrated movement, (ii) the perceived social presence of the virtual instructor, and (iii) user-experience qualities such as usability, workload, and acceptance.

To the best of our knowledge, only [9, 13, 10], proposed a contribution on this line. In [9], ARFit, a mobile AR system that lets users learn exercises through a 3D avatar and compares its effectiveness against step-by-step images and videos, showing how the avatar in AR provided the highest learning quality with respect to flat images. In the context of cycling, the authors of [13] showed that combining HMD with interactive physical tasks increases actual physical effort (higher breathing rate) and attention, and engagement, with respect to a comparable flat video and non-VR conditions. The nearest work with respect to the current contribution corresponds to [10], which implemented an AR-based yoga system that projects a life-sized 3D instructor into the user's environment, showing improved empathy, involvement, and motivation compared to traditional 2D videos. These works indicate that visualization modality may affect at least three distinct di-

mensions: exercise understanding, the social perception of the virtual instructor, and broader usability, workload, and acceptance aspects of the experience. However, they did not examine these dimensions within a unified and controlled comparison of progressively richer and comparable guidance modalities. In particular, they do not:

- Focus on general exercise learning, which often results in evaluations tied to specific tasks or contexts, limiting the generalization of findings across different exercises or (custom) training scenarios;
- Using real-world visual data for 2D exercises representations, together with non-aligned 2D/3D stimuli consistent in content and timing, may lead to instructional demonstrations that differ across modalities in subtle but relevant ways (e.g., timing, joint trajectories, execution quality). Using not aligned representations makes it difficult to isolate the effect of the visualization modality itself, as differences in movement execution may confound learning- and perception-related outcomes;
- Developing modular systems capable of supporting multiple exercises restricts scalability, as comparisons are often bound to a small set of ad-hoc examples and cannot be easily extended to broader exercise repertoires or reused under different experimental conditions;
- Systematically examining intermediate visualization modalities, which prevents understanding how different users may benefit from different levels of visual detail, embodiment, or viewpoint control when learning or recalling an exercise.

These factors limit the scalability and personalization of the experience (and the space of experimental investigations), also considering that we now have parametric human modeling that could be exploited to explore such factors [38].

We here address these gaps by introducing *XaRNold*, a modular XR fitness framework that integrates single-view, multi-view, and parametric 3D visualisations for guided workouts. *XaRNold* provides an ecologically valid tool to perform a novel comparison of (a) single-view 2D video, (b) multi-view video (in VR), and (c) an embodied 3D parametric avatar (in VR), where all exercises are perfectly aligned across conditions (for each exercise, each modality presents the same structural movements with the same poses and timing). The *XaRNold* features and adopted measurements, distinguishing our works from previous ones, are summarized in Table 1.

3 Materials and Methods

3.1 System Architecture

The system architecture of *XaRNold* is visually depicted in Figure 1, and is in line with a classical stand-alone VR application.

At its core, *XaRNold* is designed to serve different visualization modalities (in our case, single-view 2D video, multi-view 2D video, and interactive 3D

	Ours	**[10]**	**[9]**	**[13]**
Systems Features				
3D Avatar guidance	✓	✓	✓	×
Ecological validity	✓	✓	×	✓
Aligned 2D-3D content and timing	✓	✓	×	✓
Multi-View Guidance	✓	×	×	×
Data-Driven system	✓	×	×	×
3D Parametric Human Model	✓	×	×	×
Measurements				
Social presence	✓	✓	×	×
Exercise comprehension	✓	✓	×	×
Usability / Acceptance	✓	✓	✓	✓

Table 1: Comparison of main focuses vs. state of the art. Binary/ternary flags summarize whether each work covers key design and evaluation aspects. *3D avatar guidance* (exercise guidance by virtual human instructor); *Ecological validity* (evaluation in a realistic, workout-like setting and interaction flow); *Aligned 2D-3D content and timing* (identical movements and temporal cues across all visualization modalities); *Multi-view guidance* (simultaneous, synchronized multi-perspective 2D views of the same movement); *Data-driven system* (exercise motion derived from captured human data (dataset-based), as opposed to purely synthetic/hand-crafted animations); *3D Parametric Human Model* (use of a parametric body model, e.g., SMPL, to represent human shape/pose); *Social presence* (explicit measurement or design targeting perceived co-presence/companionship); *Usability / Acceptance* (explicit evaluation of system usability, cognitive demand, or technology acceptance); *Exercise comprehension* (explicit evaluation of how clearly users understand the demonstrated movement).

avatar representations) that can be rendered according to different kinds of exercises. As depicted, *XaRNold* begins with a pair of $< modality, exercises >$ user request, which queries the centralized *Exercise Knowledge Base* (EKB). EKB is required to contain temporally aligned representations of each exercise across all supported modalities, allowing consistent delivery of the same exercise through diverse visual stimuli. Once a configuration is selected, the workout scheduler retrieves the corresponding aligned exercise assets from the Exercise Knowledge Base and forwards all modality-specific data and metadata to the modular visualizer described below. In particular, the scheduler parses the configuration layer (e.g., exercise identity, repetitions, pace, viewpoint settings, and body-related metadata) and dispatches the appropriate stimulus package to the selected rendering module, whether 2D single-view video, multi-view video, or 3D avatar. The modular visualizer then renders the selected content in the virtual environment while preserving internal synchronization, so that timing, phase progression, and execution remain identical across modalities. Considering also aligned motion data, this ensures that differences are only to the visualization modality with a robust and reproducible approach.

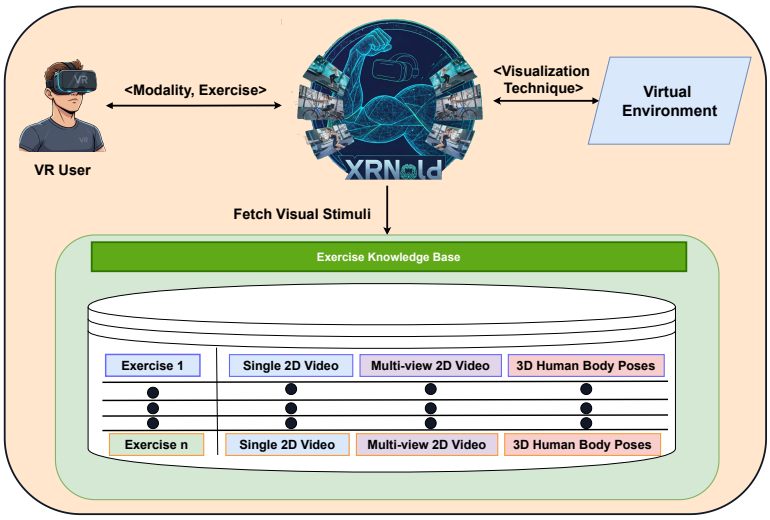


Fig. 1: System architecture of the *XaRNold* platform. Each exercise is linked to three aligned visualizations: single-view 2D video, multi-view 2D video, and 3D human-pose stimuli. The system retrieves the appropriate visual content from the Exercise Knowledge Base and delivers it to the VR environment.

3.2 System Implementation

We implemented *XaRNold* as a stand-alone VR experience. Its implementation revolves around the novel visualization modalities here introduced, namely Multi-View 2D Visualization (MV2D) and Parametric 3D Human (P3HU) (visually depicted in Figure 2).

The MV2D mode presents a synchronized four-camera display of a real human demonstrator from lateral perspectives, arranged cylindrically around the user's viewpoint. The layout is fixed in view, enabling users to observe posture and motion from multiple angles without manual navigation, following prior work on spatial visualization [35, 33], maintaining familiarity with standard video formats. The P3HU mode uses instead the SMPL parametric body model [38], animated with 3D joint trajectories extracted from real tracked humans. The avatar is rendered at life-size on a grounded virtual environment, allowing users to view the motion from any angle. This supports an embodied interaction paradigm that may foster motor resonance, imitation learning, and deeper gesture comprehension [57, 36], offering an active alternative to passive video observation. Both visualization modes are integrated into the same virtual environment and are time-locked for accurate cross-condition comparisons. This allows us to isolate user responses to different instructional representations of the same exercise content under controlled yet ecologically valid conditions.

2D Web Interface For the sake of our experimental setting, we also implemented a 2D web interface (named PC), which mimics the same logic functionali-

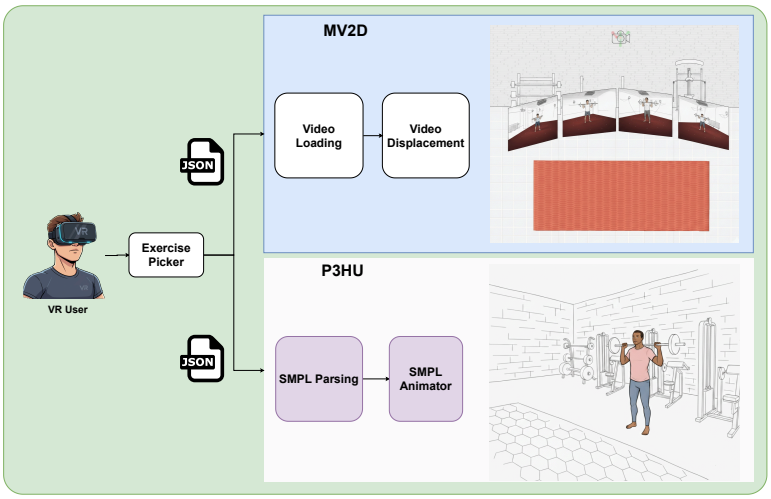


Fig. 2: Scheme of the visualization modalities presented in the *XaRNold* system. On the top, the 2D multi-view video displays a real human performing the exercise from synchronized frontal and lateral views. On the bottom, the same exercise is shown using a 3D parametric avatar, animated from corresponding motion data.

ties of MV2D and a parametric 3D model, but on a classical web interface (which mimics classical online gym systems). This is visually depicted in Figure 3.

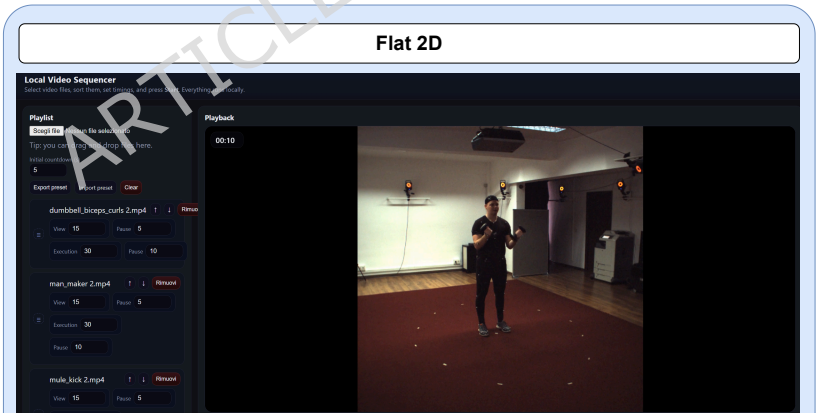


Fig. 3: Visualization of the classical 2D flat video mode.

The PC-based runs in the browser, without any server-side component (as for the VR counterpart). In this condition, the interface was displayed on a conventional desktop monitor, outside the VR headset (imitating what users do while using a classical online training system). The interface provides a video

sequencer that allows loading the exercise clips from disk, arranging them into a playlist, and associating each item with the same sequence pattern of other modalities before the participant interacts with the system. During playback, a single 2D video is shown with an overlaid countdown and phase label. This tool ensures that, in the desktop condition, participants are exposed to the same exercise content and timing structure adopted in the VR modalities, while keeping the interaction aligned with conventional screen-based fitness experiences. It is worth noticing that the PC modality serves as a natural baseline reflecting standard screen-based fitness instruction, allowing assessment of the relative benefits of immersive alternatives.

Multi-View 2D Videos Visualization In this modality, shown in Figure 4, the user can view and perform the exercise through four virtual screens arranged inside the virtual environment.

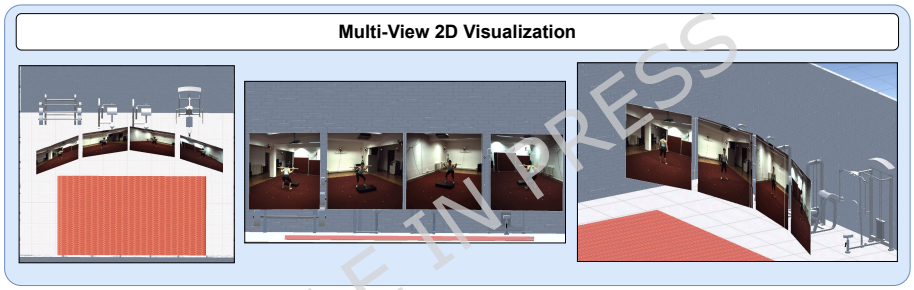


Fig. 4: Top-View and other perspectives of the *MV2D* mode for different exercises.

Each screen displays a synchronized recording of a real human demonstrator captured from a different angle: front-right, front-left, rear-right, and rear-left. These views are derived from the multi-camera setup of the Fit3D [17] dataset and are temporally aligned so that all videos show the same movement phase at the same time. This multi-angle configuration allows participants to inspect the exercise from complementary perspectives and infer depth, posture, and joint alignment more effectively with respect to a single frontal view. Such a multi-view layout provides a clearer and more technically informative depiction of the movement than a standard 2D viewpoint [17]. The four screens are embedded in the VR scene as spatially arranged panels, placed at a empirically calibrated distance of 1.7 m to ensure comfortable visualization, binocular depth perception, and rapid comparison between angles, without requiring users to walk around or reorient their whole body. At this distance, users can view two screens simultaneously and inspect the full multi-view layout with less than 90° of horizontal head rotation. Participants were then free to approach or move away by physically moving within the tracked space. The layout remains fixed with respect to the user's main field of view, so that participants can alternate between observing

and executing the exercise while maintaining a stable visual reference. In this sense, the *MV2D* modality combines the familiarity of video-based instruction with the spatial richness of a multi-camera setup, offering an intermediate step between flat video guidance and fully embodied 3D avatar visualization.

Parametric 3D Human Visualization In this modality, the user can visualize and exercise along with a virtual avatar, which performs the execution of any workout codified in our EKB. The avatar is generated using a parametric 3D human model (in our case, the SMPL) and animated with preprocessed pose data corresponding to each exercise. Different views of the same SMPL avatar performing multiple exercises are reported in Figure 5.

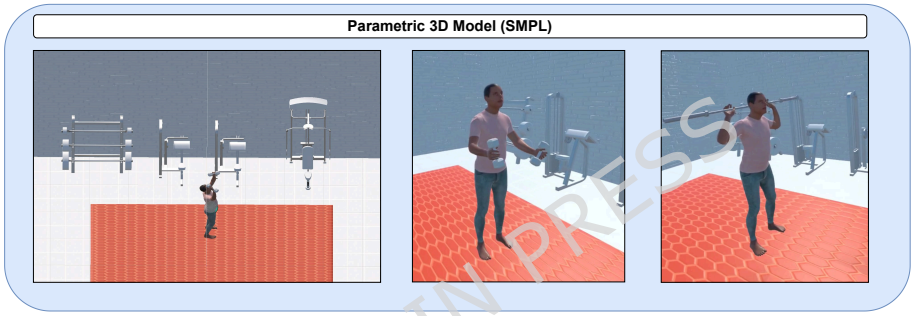


Fig. 5: Top-View and User Visualization Perspectives of the *P3HU* mode for different exercises.

This allows for a consistent and repeatable representation of movement, enabling participants to observe the motion from various perspectives in real time. Moreover, the use of a full-body avatar supports spatial embodiment and perceptual salience of joint articulation and movement transitions [51, 57]. In our implementation, the avatar is embedded in a minimal but spatially coherent VR environment. The default layout places the avatar centrally in the user's field of view. Participant height was handled automatically through the headset's built-in tracking system (via the Meta XR SDK in Unity), which continuously updates eye level and viewpoint. The avatar was instantiated using the dataset SMPL parameters, preserving the original instructor's body without rescaling, and spawned at an empirically calibrated distance of 1.5 m in front of the user, with a side-facing (profile) orientation. This allows full-body visibility at first appearance. Participants were then free to approach or move away from the avatar by physically moving within the tracked space. Participants can move within the tracked area by taking a step forward or backward, or slightly to the side, can approach or move away from the virtual trainer, and inspect the movement from different angles. It is worth noting how the *P3HU* modality provides a natural, embodied reference for imitation and motor comprehension, with a

data-driven approach: codifying an exercise means providing a list of joint-linked spatial coordinates that represent the animation of the exercise itself.

Workout scheduler We design a workout scheduler that, for each session, allows selecting exercises recorded from subjects of the chosen sex, leveraging the asset metadata that already encode body-model information (shape and appearance) together with the motion files, with an approach inspired by [3]. In particular, the scheduler works with a data-driven approach: it parses a predefined data structure (in our case as JSON) and adapts the virtual interface according to selected modality, body (and gender) and chosen exercises. In practice, adding, removing, or reordering workouts only requires specifying the list of exercises, their associated motion files, and a small set of metadata (e.g., repetitions, pace, or viewpoint configuration) in the configuration layer. This preserves a consistent structure, simplifies re-targeting across bodies, and ensures that the rendered avatar matches the participant's perceived category and visualization modality. This approach was also used to manages view/perform phases, pauses, and scene flow, operating on the sex and exercise-specific configuration

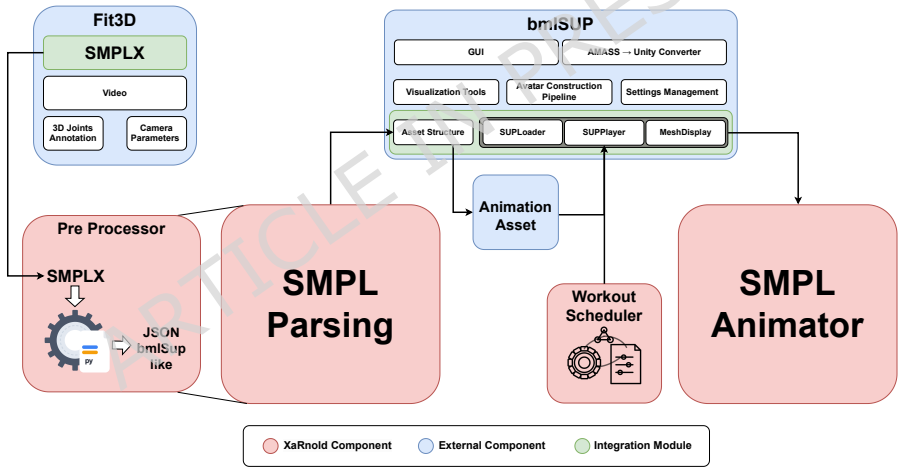


Fig. 6: Architecture of the custom SMPL-X parser implemented to project EKB annotations into an animatable SMPL Avatar in Unity.

SMPL Parser for 3D Avatar To handle motion playback in Unity, we adopted a custom parser and loader built on top of the bmlSUP plug-in [5]. Figure 6 summarizes the subset of bmlSUP modules we employ in our VR workflow and the components we omit or replace, enabling preprocessed SMPL sequences to be streamed in real time encodesnchronized with the virtual scene layout [5]. Starting from Fit3D [17], SMPL-X annotations are converted to SMPL to be

compatible with the bmlSUP engine. This procedure is achieved by a custom middleware component which aims at translating any SMPL-X annotation into the SMPL encoding structure. Starting from this parsed data, the custom developed manager component orchestrates bmlSUP components. bmlSUP was originally designed to visualize motion sequences in 2D interfaces. For this reason, we reused only its low-level components while discarding all high-level user interface and playback controls. From the parsed data, we extract all the animation assets employed by the Workout Scheduler to manage which animation is loaded, played, and rendered. These three actions are respectively implemented through the SUPLoader, SUPPlayer, and MeshDisplay. The result of this pipeline is the generation of an animated 3D SMPL avatar.

Knowledge Base The knowledge base of our system is a structured collection of exercise assets derived from the Fit3D dataset [17] that is composed of records of two human subjects performing different exercises with synchronized multi-view 2D video streams and corresponding 3D body poses obtained by fitting a parametric human model to motion-capture data. These aligned representations are the central data of our system: all three visual stimuli presented in XaRNold are rendered from the same recordings. Concerning exercise and avatar selection, for the sake of our experimental setting, we selected two subjects (male and female) from the dataset and a subset of exercises following common practices in the literature, which include six movements: *man maker*, *squat*, *mule kick*, *walk the box*, *dumbbell biceps curls*, and *warrior*. Exercises were selected based on prior literature [1, 14, 40, 14, 10]. Indeed, part of this set overlaps with movements that have already been adopted in [14, 10]. The remaining were sampled to cover both upper and lower-body movements, including actions with different amplitudes, intensities, and levels of whole-body displacement. No constraints on exercise dynamism were imposed due to the use of a head-mounted display, in order to avoid selection bias and enable a more realistic evaluation under immersive conditions. Below is a description of the selected exercises:

- Man Maker: compound full-body exercise combining push-ups, rowing motions, and an overhead press, requiring coordinated upper- and lower-body engagement;
- Squat: a lower-body exercise involving hip and knee flexion–extension, commonly used to assess fundamental lower-limb movement patterns;
- Mule Kick: exercise focuses on controlled hip extension performed in a quadrupod position, emphasizing lower-limb isolation and balance;
- Walk the box: a dynamic exercise involving coordinated stepping and arm movements in multiple directions, requiring spatial awareness and whole-body coordination;
- Dumbbell biceps curl: upper-body isolation exercise characterized by elbow flexion with minimal trunk involvement
- Warrior: pose that involves a static-to-semi-dynamic stance combining lower-body strength, balance, and upper-body posture control.

Following recommendations to treat experimental stimuli as a sample from a broader population rather than as fixed effects [26], we consider each exercise as one instance drawn from the larger class of typical instructor-led fitness movements. Our goal is to compare three visualization modalities across a reasonably diverse and realistic set of workout actions. For every exercise, we extract three types of assets, associated with a single exercise identifier in our local repository: (i) a single-view 2D clip for PC baseline condition, (ii) four synchronized camera views (front-left, front-right, rear-left, rear-right) that feed the MV2D modality, and (iii) a frame-by-frame 3D pose sequence that drives the SMPL-based avatar in the P3HU modality [38]. A key strength of this design is that the three visualization modalities are perfectly aligned at the data level: the 2D videos, the four-screen multi-view layout, and the SMPL avatar all correspond to the same physical performance by the same instructor, with matched temporal structure. As a consequence, when participants switch from one modality to another, they are always observing and reproducing the same movement; what changes is only how that movement is presented.

4 Experimental Setting

4.1 Apparatus

XaRNold was implemented using the Unity Game Engine (v 6000.0.41f1), targeting the Meta Oculus Quest 3 (OQ3) headset as the primary deployment platform (but is also compatible with the Oculus Quest 2). The system was adapted for the OQ3, using the Meta XR SDK (v78.00), along with the Meta XR Interaction SDK (v. 78.00), to support immersive rendering and interaction handling. This stack was chosen for its balance between performance, cross-platform XR flexibility, and cost-efficiency (which is key for fitness-oriented deployment) in both research and applied settings. During experimental sessions, the system ran on the headset in a completely untethered setup, granting participants full freedom of movement; this allowed us to test the solution in a "ready-to-use" form. For the tests, we adopted a ROG Strix G18 (2023) G814JI laptop, equipped with an Intel Core i9 CPU, 16 GB RAM, and an NVIDIA GeForce RTX 4070 GPU (8 GB VRAM). Pre, intra, and post-condition questionnaires were administered using the Google Forms platform. It is worth noticing that the system architecture does not require servers during runtime, as all stimuli were locally stored and dynamically loaded at runtime.

4.2 Experimental Procedure

We designed a within-subject 1×3 study conducted within a single virtual scenario, comparing three visualization modalities: **2D Single View** (baseline), **MV2D** (synchronized multi-view), and **P3HU** (3D SMPL avatar). The entire experimental process is illustrated in Figure 7.

At the beginning of the pre-questionnaire, participants provided informed consent. They then completed, in the subsequent sections of the pre-questionnaire,

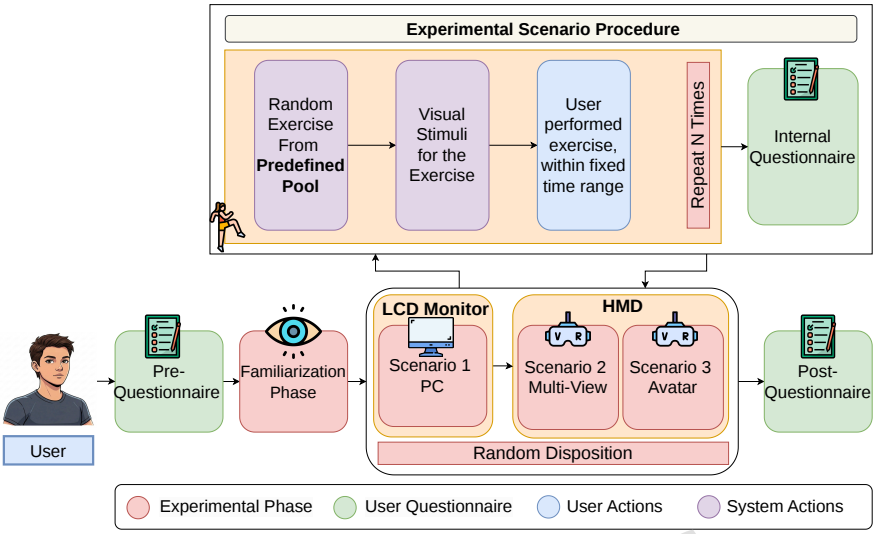


Fig. 7: Schema of the adopted experimental procedure.

items collecting demographic information (age, gender, education) and assessing familiarity with VR/XR, prior experience with immersive fitness tools, workout habits, and understanding of digital 3D representations (see Section 4.4). After providing informed consent, participants reported demographic information (age, gender, education) and completed a pre-questionnaire assessing familiarity with VR/XR, prior experience with immersive fitness tools, workout habits, and understanding of digital 3D representations (see Section 4.4). Participants then completed a familiarization phase introducing the *XaRNold* system (navigation and interpretation of visual stimuli). To avoid influencing performance in the main task, they performed a neutral warm-up (e.g., arm circles) unrelated to any of the evaluated exercises [3].

The order of the three scenarios (2D Single View, MV2D, P3HU) was randomized for each participant. In each scenario, participants completed three trials, each corresponding to a triplet of exercises sampled from the pool of six workout movements P_W (see Section 3.2).

Exercise triplets were pre-generated prior to the study following a constrained randomization strategy. Specifically, a set of 30 possible triplets was defined in advance, such that each exercise from the pool appeared the same number of times across the entire experiment. Each participant was then assigned one of these pre-generated triplets, ensuring a balanced distribution of exercises across participants and avoiding stimulus-selection bias. The same triplet of exercises was used across all three modalities. Within each scenario, however, the order of the three exercises followed a pre-generated randomized schedule. Additionally, scenario order was counterbalanced across participants, so each modality appeared equally often in first, second, and third position, mitigating sequence and

learning effects [24]. To promote user identification and avoid gender-related confounds, the instructor/avatar shown in the relevant modalities always matched the participant's sex, consistent with previous findings on avatar embodiment and virtual fitness instruction [16, 51].

Each trial consisted of an observation–execution cycle:

1. **Demonstration phase** ($T_s = 40\,\text{s}$): participants observed the movement according to the current visualization modality (frontal 2D, synchronized MV2D, or P3HU). The duration ensured that at least two complete repetitions were visible at a natural pace while limiting fatigue, in line with AR/VR fitness protocols relying on brief self-contained segments [10, 1].
2. **Preparatory pause** (10 s): a countdown allowed participants to re-center and prepare for the execution phase.
3. **Execution phase** ($T_e = 60\,\text{s}$): participants reproduced the exercise with *no visual stimulus present*. In immersive conditions, only the virtual environment remained visible. This design required reliance on an internal representation of the movement rather than real-time mirroring.
4. **Post-trial pause** (10 s): served as recovery before the next trial.

Each cycle lasted approximately 2 minutes (40 s + 10 s + 60 s + 10 s), following prior AR/VR protocols that segment the session into short, comfortable episodes [10, 1]. Removing the stimulus during the **Execution phase** dissociated the influence of the visualization modality from practice and sequence effects, encouraging performance based on an internalized representation of the movement. This approach aligns with XR and motor-learning paradigms where participants practice with rich visual feedback but are tested in blocks or trials with reduced or no feedback, typically under within-subject counterbalanced orders [18, 45, 31]. This procedure was repeated for all three exercises (N=3) in the participant-specific triplet within each modality, yielding nine executions per participant.

After each scenario, participants complete a dedicated module questionnaire. The questionnaire is administered on a PC, and participants respond only to items related to the specific modality they have just experienced. Besides collecting subjective ratings, this step also acts as a wash-out (structured break between modalities). In practice, filling in the questionnaire provides a short period in which participants can rest to reduce fatigue and mitigate positional or viewpoint-related biases before entering the next scenario, in line with general recommendations to interleave active segments with brief off-headset pauses [24]. The order of the internal questionnaires followed the scenario order, so that ratings were always provided immediately after the corresponding modality, while the experience was still fresh in participants' memory.

Upon completing all three scenarios, participants fill out a comprehensive post-questionnaire (see Section 4.3). This final survey gathers comparative feedback on the entire *XaRNold* experience, including perceived effectiveness of each visualization mode, overall preference, and future willingness to use the system in a real training context. Participants are also asked to rank the three visual-

ization modes from best to worst and to provide free-text comments to justify their top choice for qualitative analysis.

4.3 Measurements

Beyond demographics, we focused our measurement strategy on three key aspects: (i) background and prior experience of participants, (ii) modality-wise evaluation of usability, technology acceptance, workload, perceptual clarity, and social aspects, and (iii) post-experience comparative judgments on clarity, memorability, and social presence across visualization modes. Unless otherwise stated, agreement-based items were rated on a 5-point Likert scale. In the post-experience questionnaire, instead of repeating full Likert blocks for each modality, comparative questions were implemented as ranking and forced-choice items in which participants ordered the three modalities (PC, MV2D, P3HU) according to given criteria (e.g., clarity of technique, overall learning support). Moreover, the post-experience questionnaire was designed as a mixed-method analysis to capture the reasons behind participants' ratings and rankings and to identify recurring design factors not fully captured by fixed-response items. All the single-furnished items are detailed in the Supplementary Files. To make explicit how the measured constructs map into the RQs and study objectives, we here introduce Table 2, which reports their conceptual organization.

Table 2: Conceptual organization of the construct families adopted in the evaluation.

Construct family	Evaluation focus	Measures
Learning & Cognitive Outcomes (RQ1, RQ3)	Assesses how effectively each modality supports the understanding, interpretation, and perceived reproducibility of the demonstrated exercise.	**PER**, **CM**
System Quality & Technology Acceptance (RQ2)	Evaluates the perceived quality of interaction with each modality in terms of usability, workload, and acceptance as a tool for exercise support.	**SUS**, **TAM** (**PU**, **PEOU**), **NASA-TLX**
Social-Affective Perception (RQ2, RQ3)	Captures the extent to which the instructor is perceived as socially present, natural, supportive, and comfortable to follow during the exercise experience.	**SP**, **C-SOC**

Pre-questionnaire (demographics, training, and VR background). Before the experimental session, participants completed a pre-questionnaire collecting demographic information and background on physical training and XR usage. Demographic items recorded age, gender, and highest level of education. A set of items (PRE6–PRE11) assessed regular physical activity and fitness habits,

including whether participants regularly practice physical activity, the average number of training sessions per week, the main training setting (home, gym, outdoor, mixed), practice of team sports and type of sport, how they mainly learned their current exercises (e.g., in-person instructor, online videos), and whether they had a gym membership in the last 12 months. A second block (PRE12–PRE15) focused on VR/AR experience, asking participants to self-rate their XR experience level, report whether they had ever used a VR/AR HMD, specify which devices they had used (e.g., Meta/Oculus Quest, other HMDs, Cardboard, HoloLens), and indicate how often they experienced typical cybersickness symptoms (nausea, dizziness, eyestrain, headache) when using such devices. These data were primarily used to describe the sample and explore possible moderating effects of training and VR familiarity on the subsequent measures.

Internal modality-wise questionnaire

System usability (System Quality & Technology Acceptance, RQ2). For each visualization modality, we evaluated perceived usability through the System Usability Scale (SUS, 10 items), a standardized instrument widely used to gauge subjective usability of interactive systems [8, 3]. In line with common practice, we report the conventional SUS total score on the 0-100 scale as the primary summary measure [8]. However, we also retained a diagnostic decomposition of SUS into positively and negatively worded items (SUS-POS and SUS-NEG) to support interpretation, as already did in related works [18, 53]. This allowed us to distinguish between overall satisfaction (e.g., "I think that I would like to use this system frequently") and residual frustrations or perceived complexity (e.g., "I found the system unnecessarily complex"), providing a nuanced picture of usability across the three visualization modalities. In practice, this separation gave us an additional layer of usability interpretation in terms of concrete design factors, and it complements the SUS total score rather than replacing it.

Technology acceptance (System Quality & Technology Acceptance, RQ2). User acceptance of each modality was assessed using a customized version of the Technology Acceptance Model (TAM), which includes both perceived usefulness (PU-x items) and perceived ease of use (PEOU-x items) [15]. Items were adapted to explicitly frame each modality as a tool to *learn workout exercises* (e.g., "This modality helps me better understand the key aspects of the exercise technique", "Interacting with this modality is simple"). This choice was driven by the same methodology adopted in [3], considering that our interest was focused on the perceived usefulness and ease of use of each visualization mode, more than on broader technology acceptance and behavioral usage, which are more production-oriented; for this reason, Attitude Toward Using and Behavioral Intention were not assessed.

Cognitive workload (System Quality & Technology Acceptance, RQ2). Perceived mental and physical workload were measured through the classical NASA-TLX

scale [20]. We here state that we adopted the raw TLX (unweighted) version and we reverse-coded the unique positive item (i.e., NASA-TLX-4), so that all of them have the same direction. Participants rated mental demand, physical demand, temporal demand, effort, perceived performance, and frustration on 10-point scales after each condition (from 1 = *Low* to 10 = *High*), as did in [18, 53]. This allowed us to test whether immersive visualization (MV2D and P3HU) imposed additional cognitive or physical demands compared to the PC baseline and to relate usability and acceptance scores to the subjective cost of using each modality.

Perceptual effectiveness of the instructional content (Learning & Cognitive Outcomes, RQ1, RQ3). While SUS, TAM, and NASA-TLX provide valuable information about overall usability, acceptance, and workload, they do not directly capture *how well* users can visually parse, understand, and remember the demonstrated movements. Standard instruments generally treat the system as a generic interactive interface, without explicitly modelling key affordances of visual exercise instruction such as clarity of joint articulation, depth perception, or support for self-correction. To address this, we introduced a custom construct to assess how effectively each modality supports the perception and understanding of the movement (PER-x items). The PER block included 13 items evaluating: (i) visual clarity (e.g., visibility of critical angles and range of motion), (ii) spatial understanding (e.g., depth, joint alignment), (iii) temporal guidance (e.g., ability to follow rhythm and timing), (iv) perceived error detection and self-correction ability, and (v) confidence in task execution and memory retention. Example items include "The modality made joint alignments clear (e.g., knee-foot, back-pelvis)" and "I feel able to correct myself autonomously when using this modality". All the selected items for perceptual effectiveness were derived and adapted from validated scales in the literature on XVR yoga, exergames, and motion-based training systems [10, 60, 1, 32, 12, 25, 56]. The modality-wise questionnaire thus aims at evaluating how well each visualization modality supports users in understanding the exercise, together with its usability and workload, anticipating potential benefits for long-term exercise comprehension.

Social presence and perception of the instructor (Social-Affective Perception, RQ2, RQ3). To capture the social dimension of guidance in each condition, we included a dedicated Social Presence construct (SP-x items) within the internal, modality-wise questionnaire. After each scenario, participants responded to the same seven statements referring to the "instructor", intended as the person shown in the 2D or MV2D videos in the PC-based conditions, and as the Avatar in the P3HU condition. The scale evaluated the perceived social presence, comfort and spatial proximity, and relatability of this instructor across all three visualization modes. Items covered three main facets: (i) *co-presence and companionship* (e.g., feeling that the instructor was "there" exercising together with the participant, rather than being a purely instrumental visual reference: "I perceived the instructor's social presence alongside me"), (ii) *comfort and spatial appropriateness* (e.g., adequacy of distance, size, and position in the field

of view: "The instructor's distance and scale were appropriate for learning"), and (iii) *relatability and naturalness* (e.g., feeling at ease with the instructor and perceiving movements as natural rather than mechanical: "I felt comfortable interacting with the instructor"). The wording of these items was informed by adaptations of social presence questionnaires used to evaluate social involvement with an instructor's lessons [10] and by prior work on how embodiment and avatar design influence social and affective responses in VR [14, 16]. Following the rationale adopted in the yoga study [10], we emphasized empathy and behavioural coupling (e.g., the extent to which one's actions depend on the instructor's actions). Reliability analyses (Section 5.2) confirmed acceptable internal consistency for this construct, and SP scores were computed separately for each modality, yielding a social presence profile for PC, MV2D, and P3HU conditions.

Post-experience questionnaire

Comparative clarity and preference (Learning & Cognitive Outcomes, RQ1, RQ3). Following the completion of all experimental conditions, participants were asked to complete a comprehensive post-experience questionnaire aimed at comparing the three visualization modalities in terms of clarity and preference. In the initial design, we considered administering additional 5-point Likert scales per modality; however, this would have led to substantial redundancy with the modality-wise constructs (SUS, TAM, NASA-TLX, PERC) and to an excessively long questionnaire, leading to reduced efficiency and diminished reliability of the responses. Moreover, previous work on subjective assessment and user experience has shown that, in within-subject designs where participants evaluate multiple alternatives of the same underlying content, comparative judgments such as pairwise choices, rankings, or relative preferences can yield more discriminative and internally consistent data than repeated absolute ratings on Likert scales [61, 11, 42]. For this reason, the comparative section was implemented using a more compact set of forced-choice and ranking items, focusing on the relative ordering of the three modalities rather than on repeated absolute ratings.

The first section consisted of a comparative modality assessment block, where participants expressed their preferences and comparative judgments across the PC-based, MV2D, and P3HU conditions. These items captured participants' perceived ability to understand the exercise technique from each modality (CM-x items) by interpreting visual information to replicate the gesture with confidence. In line with comparative UX research that favors comparative judgments over repeated absolute ratings to reduce scale–use variance [61], and with VR quality og experience studies that adopt rank-based protocols for immersive content [11, 42], we collected rank-order preferences across the three modalities (PC, MV2D, P3HU) rather than full pairwise comparisons. Specifically, CM items asked participants to *classify* the modalities from best to worst for clarity-related aspects (e.g., "clearest visual information for learning", "easiest to follow"), while FC items used single *forced-choice* prompts to select the overall best (e.g., "Which

modality worked best for you?" / "Which was clearest to learn the technique?"). This design preserves the benefits of comparative judgments without the cognitive load of exhaustive pairwise comparisons, and matches the structure of our post-experience questionnaire.

Additionally, participants were asked to indicate their preferred modality overall and their willingness to use it in future training (FC-x items), as well as to provide a rank for the three modalities in terms of overall instructional support. Together, CM and FC provide a direct and user-centric summary of which visualization is perceived as the most effective teaching medium, complementing the condition-wise Likert constructs with an explicitly comparative perspective, while following recommendations to use compact comparative protocols to keep VR studies manageable in duration and cognitive load [11].

Comparative social experience across modalities (Social-Affective Perception, RQ2, RQ3). The second section of the post-experience questionnaire focused on perceived social interaction (C-SOC-x items), including which modality made the instructor or guide feel most socially present and appropriate. Participants compared the three modalities in terms of perceived co-presence, sense of being accompanied during the exercise, and appropriateness of the instructor's distance and scale. The wording of these comparative items was directly inspired by the same social presence formulations used for the SP construct, where social involvement with a remote teacher is explicitly assessed [10]. At the same time, our C-SOC items were framed at a *relative* level (e.g., "In which modality did you feel most accompanied by the instructor?") to directly compare the three visualization modes on these social dimensions, in line with broader conceptualizations of social presence in XR that emphasize comparative judgments across media configurations [44]. As with CM and FC, we opted for concise comparative questions and ranking items instead of full Likert scales for each modality, to avoid excessive duplication of social items already present in the modality-wise Social Presence construct (SP) while still capturing the relative ordering of modalities from a social perspective. In this sense, the C-SOC block reuses the same underlying social presence dimensions defined for SP, but translates them into a compact, comparative format motivated by the considerations on ranking and forced-choice protocols discussed above for CM and FC.

Qualitative Items. Qualitative data were collected to explain the mechanisms behind observed quantitative differences across modalities and to provide design implications. In particular, we collected open-ended feedback (think-aloud and free-text) to complement the quantitative findings at the end of the experimentation (and so after trying all three experimental modes). This has the goal of clarifying why specific modalities were perceived as better/worse, and to provide concrete user-centered design implications that cannot be fully captured by quantitative analysis alone. At the same time, this allowed us to outline explicit limitations of the XaRNold system, laying the basis for future advancements.

4.4 Participants

In our experimental setting, we recruited 30 participants ($M_{age} = 27.33$, $SD = 4.58$), most of whom were based in central Italy. Participants were recruited through convenience sampling from both the university environment and the local community. Recruitment was conducted via university channels, including faculty-shared announcements, in-class invitations, and on-campus outreach, as well as through word-of-mouth dissemination across personal and academic networks. Participation was voluntary.

The sample was mostly gender-balanced (43.5% female and 56.5% male). In terms of educational background, the sample was well-educated: 14 held a Master's degree, 7 a Bachelor's degree, 6 a Doctoral degree, and 3 a high school diploma or equivalent. Most participants reported being physically active, with 23 regularly engaging in physical activity and 18 having held a gym membership in the past 12 months. On average, participants trained 2–3 times per week, primarily in gyms, at home, or outdoors, with a few reporting mixed settings. Regarding exercise learning sources, most participants reported learning through in-person instruction in either individual (personal training; 8) or group-based formats (structured classes; 6), followed by self-guided methods such as online videos (3) or fitness apps (1).

Self-rated exercise experience averaged 3.35 on a 5-point scale. In terms of XR exposure, 18 participants had previously used a VR/AR headset, predominantly Meta/Oculus Quest devices, with some variety of other devices, including smartphone-based headsets, HTC Vive, and PlayStation VR. While most users reported mild or no VR-related discomfort (11/18), a slightly smaller subset experienced symptoms such as eyestrain or dizziness (7/18).

4.5 Ethics

Written consent to participate in this experimental study was collected from each subject. The entire experimental session was possible thanks to the protocol adopted in September 2025 by the *Comitato Etico della Ricerca di Ateneo* of the University of Macerata (Ethical Committee for Research, Protocol n. 117651, 17 September 2025), granting consent to carry out these experiments. This protocol allowed participants to safely employ the devices available in our laboratory (VRAI Lab). Each participant went to the Lab location and participated in a single session, which lasted, on average, 40 minutes.

5 Results

5.1 Statistical Analysis Framework

In this section, we present the rationale behind our statistical analysis. First, we assessed the reliability of our questionnaire constructs through Cronbach's alpha. The recommended threshold of $\alpha \geq 0.7$, indicates acceptable internal consistency and construct reliability [55]. We applied this to all the constructs

included in the within-subject internal questionnaire.

Then, to answer our RQs, we structured our analysis towards two main processes: the **Visualization Modality Comparison** and the **Comparative Perception Analysis**. The former examines how user responses to each construct differ across the three visualization conditions: *PC*, *MV2D*, and *P3HU*. This comparison aims to evaluate whether the immersive and embodied nature of 3D avatar guidance provides measurable benefits over conventional 2D video-based representations (even multi-view). The latter focuses on measuring a final and direct comparison across the three modalities through a ranking approach. Finally, we included a qualitative analysis, summarizing user comments collected through our think-aloud approach.

Considering the **Visualization Modality Comparison**, given the large number of measured dimensions and the heterogeneity of the investigated constructs, we adopted a two-layer analysis strategy, aiming at providing a diagnostic understanding of why one modality differs from another. We so report (i) construct-level scores as descriptive outcomes (i.e., aggregated scores computed by averaging the items belonging to the same construct), and (ii) item-level statistical analysis as a fine-grained explanatory layer that highlights which specific facets of a construct drive observed differences. This choice was motivated by our scope (i.e., comparing visual guidance interfaces for fitness, to discover fine-grained preference leading factors) and inspired by previous works [18, 53, 10].

Considering (ii) we first assessed the variable distributions using the Shapiro–Wilk test [48]. We here anticipate that, since all variables did not conform to normality, we adopted the non-parametric Wilcoxon signed-rank test [49] for pairwise within-subject comparisons across modalities. This test was chosen because it identifies significant differences between two non-parametric distributions. The results of the test (two-tailed) are first visually presented through annotated boxplots using the following significance coding: * ($1.00 \times 10^{-2} < p \leq 5.00 \times 10^{-2}$), ** ($1.00 \times 10^{-3} < p \leq 1.00 \times 10^{-2}$), *** ($1.00 \times 10^{-4} < p \leq 1.00 \times 10^{-3}$), and **** ($p \leq 1.00 \times 10^{-4}$).

When significant differences were found (in the two-tailed test), we identify the direction (e.g., higher/lower for one modality), through the matched-pairs rank-biserial correlation (RBC) as effect size (r), a non-parametric measure of association quantifying the degree to which one condition tends to yield higher scores than the other across participants [29]. If no statistical significance was observed, the comparison was retained in descriptive form only. We here state that for all our analyses, we set $p \leq 0.05$.

While discussing the obtained results, considering that item-level testing entails a large number of pairwise comparisons (which increases the probability of Type-I errors if unadjusted p-values are interpreted at face value), we applied a test-correction employing False Discovery Rate (FDR) across each family of item-level Wilcoxon tests using the two-stage Benjamini-Hochberg [6, 46]. We report FDR-adjusted p-values (p_{FDR}) and base significance decisions on the corrected values.

Regarding the **Comparative Perception Analysis**, we directly compare the ranking given by our participants, through a data-visualization-supported anal-

ysis.
We here state that all our analyses were conducted in Python 3.10 using the following library versions: pandas (v 2.3.2), numpy (V 2.2.6), matplotlib (v 3.10.6), seaborn (v 0.13.2), scipy (v 1.15.3), pingouin (v 0.5.5), statannotations (v 0.7.2), statsmodels (v 0.14.5), and pingouin (v 0.5.5).

5.2 Reliability Analysis

The results obtained by applying Cronbach's α test are reported in Table 3, which exhibits reliable scores for all of the constructs. Although the obtained results are aligned with the literature for the well-established SUS, TAM, and NASA constructs, it is worth mentioning that: (i) our custom PER scale, designed to capture gesture perception and self-correction ability, shows acceptable reliability ($\alpha \approx 0.76$); (ii) the social presence scale achieves a high reliability level ($\alpha \approx 0.85$), indicating that perceptions of the instructor's presence, comfort, identification, and behavioral naturalness are measured consistently across items.

Construct	**Cronbach's α**	**95% CI**
PER	0.765	[0.683, 0.833]
SP	0.850	[0.794, 0.894]
TAM-PU	0.916	[0.884, 0.940]
TAM-PEOU	0.829	[0.765, 0.879]
SUS-SCORE*	0.807	[0.742, 0.861]
SUS-POS	0.778	[0.696, 0.843]
SUS-NEG	0.743	[0.649, 0.819]
NASA-TLX*	0.751	[0.662, 0.823]

Table 3: Cronbach's alpha values and confidence intervals for each construct (*negative items were reverse-coded).

5.3 Visualization Modality Comparison

To provide first insights on how our users perceived the different modalities, we here report aggregated statistics for each main construct, included in our internal questionnaire, in Table 4. The results indicates a consistent advantage of the P3HU modality across most constructs: users reported higher social presence (SP), better perception (PER), higher perceived usefulness (TAM-PU), and higher satisfaction (SUS-POS) when interacting through the avatar-based interface. However, the SUS-score outlined a negligibly higher usability with respect

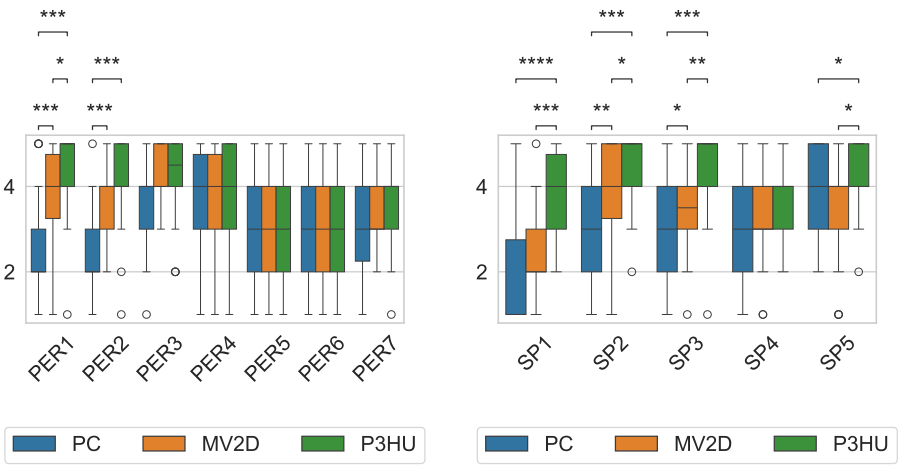
to the P3HU, in particular concerning SUS-NEG items. The same phenomenon was observed for the TAM-PEOU construct. The P3HU condition also shows lower workload (NASA-TLX), reinforcing that users perceived the interaction as less demanding. However, it is worth noticing that the MV2D condition generally outperforms PC, particularly in SP, PER, TAM-PU, and SUS-POS, indicating that immersive visualization (even without a full avatar) provides value over traditional desktop visualization and interaction. To confirm these descriptive insights, we now proceed to report statistical evidence, with the approach described in Section 5.1.

Construct	PC	MV2D	P3HU
PER (↑)	3.110 (±1.25)	3.476 (±1.16)	**3.710 (±1.22)**
SP (↑)	2.860 (±1.36)	3.287 (±1.24)	**4.053 (±0.96)**
TAM-PEOU (↑)	**4.360 (±1.00)**	4.040 (±0.89)	4.280 (±0.94)
TAM-PU (↑)	2.813 (±1.07)	3.700 (±1.07)	**4.013 (±1.03)**
SUS-SCORE (↑)	**76.83 (±17.26)**	68.67 (±14.85)	76.08 (±13.99)
SUS-NEG (↓)	**1.627 (±0.95)**	2.200 (±1.15)	1.940 (±0.94)
SUS-POS (↑)	3.773 (±1.26)	3.693 (±1.00)	**4.027 (±0.95)**
NASA-TLX (↓)	4.000 (±2.43)	4.256 (±2.32)	**3.506 (±1.96)**

Table 4: Descriptive statistics (mean and standard deviation) for each construct across the three visualization modalities. Best results are **bolded**, while second-best results are underlined.

Exercise Perception As illustrated in Figure 8(a), statistically significant differences emerged for the first three perception-related items (PER-x), favoring both MV2D and P3HU conditions, also when considering both uncorrected p-values and FDR-corrected significance (p_{FDR}). For **PER1** (viewpoint clarity), both immersive modalities were significantly more effective than PC: PC vs MV2D and PC vs P3HU were significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 28.0$ and a large effect ($r = 0.797$) for PC vs MV2D, and $W = 31.0$ with a large effect ($r = 0.847$) for PC vs P3HU. In addition, MV2D vs P3HU was significant both before and after correction ($p < 0.05$ and $p_{FDR} < 0.05$), with $W = 54.5$ and a medium effect ($r = 0.528$), indicating that P3HU provided clearer and more exploitable viewpoints than MV2D.

For **PER2** (joint alignment clarity), both immersive modalities again outperformed PC robustly: PC vs MV2D ($W = 19.0$, $r = 0.850$) and PC vs P3HU ($W = 15.0$, $r = 0.915$) were significant both before and after correction (both $p < 0.05$ and $p_{FDR} < 0.05$). In contrast, MV2D vs P3HU did not reach signifi-



(a) PER (perceptual clarity) scores.

(b) SP (social presence) results.

Fig. 8: Boxplots with annotation of statistically significant difference in two-tailed Wilcoxon-test. Reported statistical analyses are FDR-corrected.

cance (both $p \geq 0.05$ and $p_{FDR} \geq 0.05$), suggesting that both immersive formats similarly improved perceived joint alignment clarity relative to PC.

For **PER3** (gesture boundaries), PC vs MV2D and PC vs P3HU showed nominal uncorrected differences ($p < 0.05$) with $W = 35.0$ ($r = 0.591$) and $W = 63.0$ ($r = 0.502$), respectively, indicating a tendency for clearer initial/final position understanding in immersive formats. However, both effects did not remain significant after applying FDR correction ($p_{FDR} \geq 0.05$); therefore, we interpret them as trends only. No difference emerged between MV2D and P3HU (both $p \geq 0.05$ and $p_{FDR} \geq 0.05$).

Items **PER4** through **PER7** did not show significant differences in either uncorrected or corrected analyses (all $p \geq 0.05$ and all $p_{FDR} \geq 0.05$), indicating comparable perceptions of visual confusion, self-monitoring, autonomous correction, and confidence in repeating the gesture across visualization modes.

Overall, these results indicate that PC provides the weakest support for perceiving viewpoint-relevant details and joint alignment cues. MV2D improves perceptual clarity by adding complementary perspectives, but its benefit is more limited than P3HU for viewpoint clarity (PER1). P3HU yields the clearest perception of movement, consistent with a life-scale representation that supports viewpoint inspection and reduces ambiguity in critical visual cues.

Social Presence Figure 8(b) reports comparisons for items evaluating the perceived social presence of the virtual instructor. Significant results emerged for four out of five items. Significant differences emerged on Social Presence (SP) items when considering both uncorrected p-values and FDR-corrected significance (p_{FDR}).

For **SP1** (co-presence), P3HU was rated significantly higher than both PC and MV2D: PC vs P3HU and MV2D vs P3HU were significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 15.0$ and a large effect ($r = 0.921$) for PC vs P3HU, and $W = 26.0$ with a large effect ($r = 0.840$) for MV2D vs P3HU. In contrast, PC vs MV2D showed a nominal uncorrected trend ($p < 0.1$) with $W = 85.5$ and a moderate effect ($r = 0.430$), but it did not reach significance (and therefore is not interpreted as a reliable difference).
For **SP2** (distance/scale appropriateness), both immersive modalities were perceived as significantly more appropriate than PC, and P3HU was also preferred over MV2D. Specifically, PC vs MV2D ($W = 15.5$, $r = 0.819$) and PC vs P3HU ($W = 11.0$, $r = 0.932$) were significant both before and after correction ($p < 0.05$ and $p_{FDR} < 0.05$). Moreover, MV2D vs P3HU was significant both before and after correction ($p < 0.05$ and $p_{FDR} < 0.05$), with $W = 16.5$ and a medium-to-large effect ($r = 0.725$), indicating that the instructor's scale/distance was perceived as most appropriate in P3HU.
For **SP3** (comfort with the instructor), both immersive modalities outperformed PC, and P3HU further outperformed MV2D. PC vs MV2D ($W = 26.0$, $r = 0.660$), PC vs P3HU ($W = 14.0$, $r = 0.907$), and MV2D vs P3HU ($W = 18.5$, $r = 0.805$) were all significant both in the uncorrected analysis ($p < 0.05$) and after correction ($p_{FDR} < 0.05$).
For **SP4** (relatability/identification), no statistically significant effects were observed in either uncorrected or corrected analyses (all $p \geq 0.05$ and all $p_{FDR} \geq 0.05$), indicating comparable ratings across modalities regarding identification with the instructor.
For **SP5** (naturalness/supportiveness), P3HU was rated significantly higher than both PC and MV2D: PC vs P3HU and MV2D vs P3HU were significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 48.0$ and a medium effect ($r = 0.584$) for PC vs P3HU, and $W = 38.5$ with a medium-to-large effect ($r = 0.667$) for MV2D vs P3HU. In contrast, PC vs MV2D did not show a significant difference (both $p \geq 0.05$ and $p_{FDR} \geq 0.05$).
Overall, PC yields the lowest sense of co-presence and companionship, MV2D increases social comfort and perceived appropriateness of instructor placement relative to PC, and P3HU is consistently perceived as the strongest modality for social presence. This pattern is consistent with life-scale embodiment and viewpoint affordances in P3HU, which can foster a stronger sense of "being with" the instructor than fixed 2D video layouts.

Usability, Cognitive Load and Technology Acceptance We begin considering factors detailed in the NASA-TLS, TAM, and SUS scale. Figure 9 reports visually annotated statistical differences that emerged in the former, across modalities.

Significant differences emerged on NASA-TLX items when considering both uncorrected p-values and FDR-corrected significance (p_{FDR}).

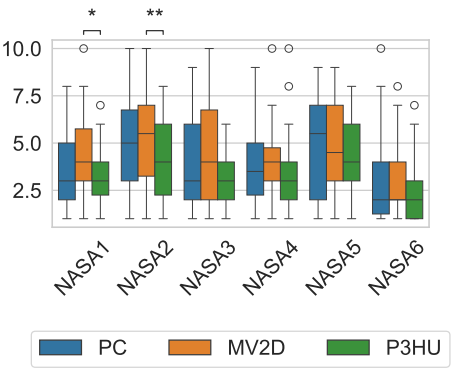
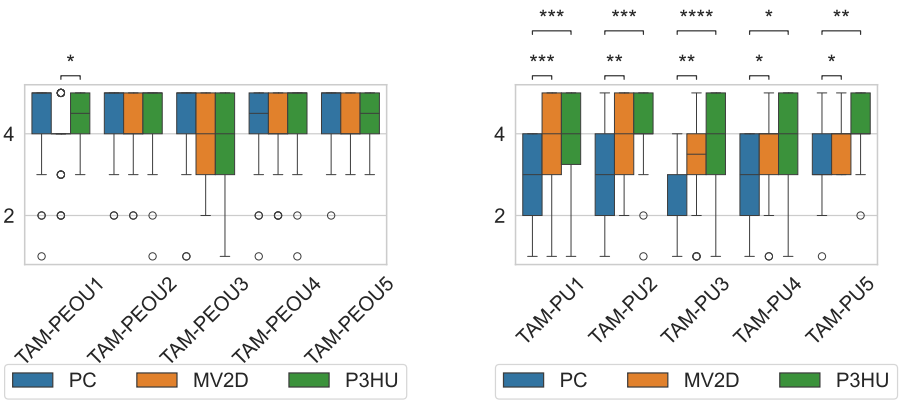


Fig. 9: NASA-TLX analysis annotated boxplots. Reported statistical analyses are FDR-corrected.

For **NASA1** (Mental Demand), MV2D vs P3HU was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 41.0$ and a medium-to-large effect ($r = 0.610$), indicating that participants experienced lower cognitive load in P3HU than in MV2D. The remaining NASA1 comparisons were not significant after correction ($p_{FDR} \geq 0.05$).
For **NASA2** (Physical Demand), MV2D vs P3HU was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 9.0$ and a large effect ($r = 0.850$), indicating that participants perceived P3HU as less physically demanding than MV2D. In contrast, PC vs P3HU showed only a nominal uncorrected difference ($p < 0.05$) with $W = 43.0$ and a medium-to-large effect ($r = 0.547$), but it did not remain significant after applying FDR correction ($p_{FDR} \geq 0.05$); therefore, we interpret it as a trend only. The PC vs MV2D comparison was not significant (both $p \geq 0.05$ and $p_{FDR} \geq 0.05$).
For **NASA3** (Temporal Demand), MV2D vs P3HU showed a nominal uncorrected difference ($p < 0.05$) with $W = 79.0$ and a moderate effect ($r = 0.473$), suggesting that MV2D was perceived as more rushed than P3HU. However, this effect did not remain significant after applying FDR correction ($p_{FDR} \geq 0.05$); therefore, we interpret it as a trend only. The remaining NASA3 comparisons were not significant after correction ($p_{FDR} \geq 0.05$).
No statistically significant effects were observed for the other comparisons and for **NASA4**, **NASA5**, and **NASA6** in either the uncorrected or corrected analyses (all $p \geq 0.05$ and all $p_{FDR} \geq 0.05$), indicating that perceived task success, effort, and frustration were comparable across visualization modes.
Regarding cognitive load, the preference tends towards P3HU (as also indicated in Table 4 but being statistically valid only for a couple of items.

Figure 10 reports statistically significant comparisons for both **TAM-PEOU** and **TAM-PU** constructs.



(a) TAM-PEOU (ease of use). (b) TAM-PU (perceived usefulness).

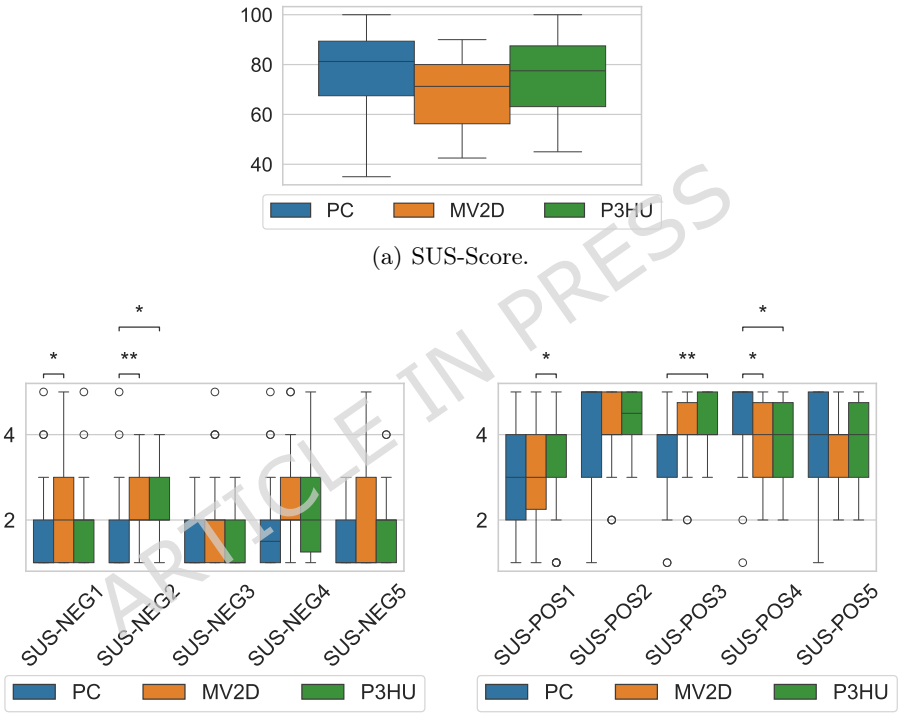
Fig. 10: Item-level comparisons for (a) TAM-PEOU and (b) TAM-PU constructs across PC, P3HU, and MV2D conditions. Reported statistical analyses are FDR-corrected.

Significant differences emerged on TAM constructs when considering both uncorrected p-values and FDR-corrected significance (p_{FDR}).

For **TAM-PEOU** (Figure 10(a)), only **TAM-PEOU1** revealed a significant difference between the two VR modalities: MV2D vs P3HU was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 24.0$ and a medium-to-large effect ($r = 0.686$), indicating that P3HU was perceived as simpler to interact with than MV2D. In contrast, the remaining PEOU items did not show significant differences after correction (all $p_{FDR} \geq 0.05$), including comparisons involving the PC baseline. Notably, **TAM-PEOU2** (PC vs MV2D) showed a nominal uncorrected effect ($p < 0.05$) with $W = 34.0$ and a medium-to-large effect ($r = 0.556$), but it did not remain significant after applying FDR correction ($p_{FDR} \geq 0.05$); therefore, we interpret it as a trend only. Similarly, **TAM-PEOU5** (PC vs MV2D) did not reach uncorrected significance ($p \geq 0.05$) and was not significant after correction ($p_{FDR} \geq 0.05$).

For **TAM-PU** (Figure 10(b)), immersive modalities consistently outperformed the PC baseline across all items. For **TAM-PU1**, both PC vs MV2D ($W = 31.5$, $r = 0.790$) and PC vs P3HU ($W = 37.5$, $r = 0.802$) were significant both before and after correction ($p < 0.05$ and $p_{FDR} < 0.05$), indicating higher perceived usefulness of immersive modalities for understanding key aspects of the exercise technique. Similarly, **TAM-PU2** and **TAM-PU3** showed significant advantages of both MV2D and P3HU over PC (all $p_{FDR} < 0.05$), with large effects especially for PC vs P3HU (TAM-PU2: $W = 14.0$, $r = 0.907$; TAM-PU3: $W = 10.0$, $r = 0.938$). **TAM-PU4** and **TAM-PU5** also confirmed significant benefits of both immersive formats over PC after correction (all $p_{FDR} < 0.05$), with medium-to-large effects (TAM-PU4: PC vs MV2D $W = 46.0$, $r = 0.562$; PC

vs P3HU $W = 41.0$, $r = 0.676$; TAM-PU5: PC vs MV2D $W = 47.0$, $r = 0.552$; PC vs P3HU $W = 37.5$, $r = 0.750$).
In contrast, no statistically significant differences were observed between MV2D and P3HU for any TAM-PU item after correction (all $p_{FDR} \geq 0.05$). While some MV2D vs P3HU comparisons showed nominal uncorrected tendencies (e.g., **TAM-PU3** and **TAM-PU5**, both $0.05 < p < 0.1$), these did not meet the corrected significance criterion and are therefore interpreted as trends only.
These results suggest that both immersive modalities improved perceived usefulness relative to the PC baseline, whereas their relative differences tend in favor of P3HU.



(a) SUS-Score.

(b) SUS-Negative items.

(c) SUS-Positive items.

Fig. 11: SUS item-level comparisons across PC, P3HU, and MV2D modalities. Questions were indexed according to their index as Negative or Positive answers. Reported statistical analyses are FDR-corrected.

Figure 11(a) reports the SUS-score according to modality. As reported, PC and P3HU obtained comparable overall usability scores, while MV2D showed a lower descriptive tendency (which is below the commonly adopted usability acceptability threshold, which is 70 [8]). However, no statistically significant differences emerged among the three modalities. Therefore, we proceeded to

item-level analysis to identify which specific usability aspects contributed to the observed pattern. On this line, Figure 11(b) and Figure 11(c) present the SUS item-level analysis. Significant differences emerged on SUS items when considering both uncorrected p-values and FDR-corrected significance (p_{FDR}). For the **SUS-NEG** items (Figure 11(b)), statistically significant differences were observed mainly for perceived complexity and required support. **SUS-NEG1** (system unnecessarily complex) showed that PC vs MV2D was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 32.5$ and a medium-to-large effect ($r = 0.658$), indicating higher perceived complexity in MV2D than in PC. In contrast, MV2D vs P3HU showed only a nominal uncorrected difference ($p < 0.05$) with $W = 53.5$ and a moderate effect ($r = 0.490$), but it did not remain significant after correction ($p_{FDR} \geq 0.05$); therefore, we interpret it as a trend only.

SUS-NEG2 (need support of a technical person) confirmed a robust familiarity advantage for the PC baseline: both PC vs MV2D ($W = 40.0$, $r = 0.710$) and PC vs P3HU ($W = 75.5$, $r = 0.535$) were significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), indicating that participants felt they would require less external support in PC than in both immersive modalities (outlining MV2D as the worst). No difference was observed between MV2D and P3HU (both $p \geq 0.05$ and $p_{FDR} \geq 0.05$).

For the remaining negative items, no effects remained significant after correction (all $p_{FDR} \geq 0.05$). Notably, **SUS-NEG4** (cumbersome to use) showed a nominal uncorrected difference for PC vs MV2D ($p < 0.05$) with $W = 40.0$ and a medium-to-large effect ($r = 0.579$), but it did not remain significant after FDR correction ($p_{FDR} \geq 0.05$), and is therefore interpreted as a trend only. Similarly, **SUS-NEG5** (needed to learn a lot before getting going) showed only a nominal uncorrected difference for PC vs P3HU ($p < 0.05$) with $W = 39.0$ and a medium-to-large effect ($r = 0.544$), but it did not remain significant after correction ($p_{FDR} \geq 0.05$). **SUS-NEG3** (inconsistency) did not show significant differences in either uncorrected or corrected analyses (all $p \geq 0.05$ and all $p_{FDR} \geq 0.05$).

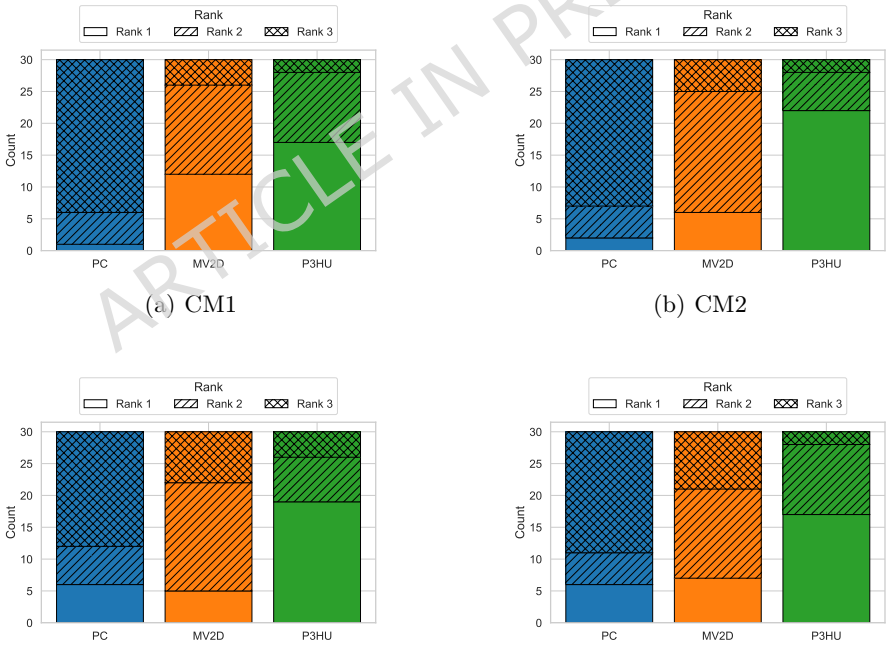
For the **SUS-POS** items (Figure 11(c)), fewer robust differences were observed after correction, but they highlight complementary strengths of P3HU and PC. **SUS-POS1** (like to use frequently) revealed a significant difference between the two VR modalities: MV2D vs P3HU was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 28.0$ and a large effect ($r = 0.634$), indicating higher willingness to use P3HU frequently compared to MV2D.

SUS-POS3 (functions well integrated) showed that PC vs P3HU was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$), with $W = 18.0$ and a large effect ($r = 0.789$), indicating higher perceived integration for P3HU than PC.

SUS-POS4 (most people would learn quickly) highlighted a learning advantage for the PC baseline over MV2D: PC vs MV2D was significant in the uncorrected analysis ($p < 0.05$) and remained significant after correction ($p_{FDR} < 0.05$),

with $W = 52.0$ and a medium-to-large effect ($r = 0.589$). In contrast, PC vs P3HU showed a nominal uncorrected difference ($p < 0.05$) with $W = 49.5$ and a medium effect ($r = 0.529$), which remain significant after correction ($p_{FDR} \leq 0.05$).
All remaining positive items did not show significant effects after correction (all $p_{FDR} \geq 0.05$). In particular, **SUS-POS2** (easy to use) showed a nominal uncorrected difference for MV2D vs P3HU ($p < 0.05$) with $W = 30.0$ and a medium-to-large effect ($r = 0.608$), but it did not remain significant after FDR correction ($p_{FDR} \geq 0.05$); thus, it is interpreted as a trend. **SUS-POS5** (confidence using the system) did not show significant differences.
To summarize, P3HU was mostly perceived as a useful modality with respect to the others (TAM) while yielding lower or comparable cognitive and physical demands (NASA). By contrast, the PC modality was consistently rated as the easiest to use, with users tending to prefer P3HU over MV2D.

5.4 Comparative Perception Analysis



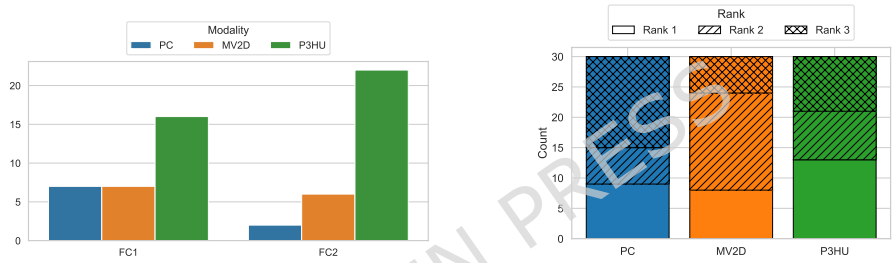
(a) CM1 (b) CM2

(c) CM3 (d) CM4

Fig. 12: Item-level comparisons for CM-X items constructs across PC, P3HU, and MV2D conditions. Rank-x indicates the number of participants who placed a modality in x-place ($x \in \{1, 2, 3\}$); stacked bars show ranking frequency.

We here delineate the analysis of our post-questionnaire results, which exploits CM, FC, and C-SOC constructs to directly assess which modality was perceived as best. Figure 12 illustrates CM-x items ranking across the three modalities.

In terms of the four core instructional clarity dimensions, gesture comprehension (CM1), visual information quality (CM2), boundary identification (CM3), and followability (CM4). Across all constructs, the P3HU modality emerged as the top-ranked condition, receiving the highest number of *rank-1* votes. In contrast, the PC modality was most frequently ranked last. Regarding the FC-x items, the comparative user feedback (Figure 13) reinforces the perceived instructional superiority of the P3HU modality.

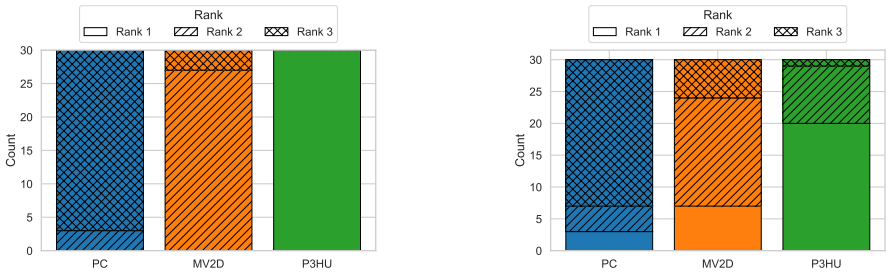


(a) FC-1 and FC-2. (b) FC-3.

Fig. 13: Item-level comparisons for FC-X items constructs across PC, P3HU, and MV2D conditions. Chart (a) reports counts for FC-1 (preferred modality) and FC-2 (perceived clarity of gesture execution) across modality. Chart (b) reports FC3 as a rank-based comparison: rank-x indicates the number of participants who placed a modality in x-place ($x \in \{1, 2, 3\}$); stacked bars show ranking frequency.

In terms of both **preferred modality** (FC1) and **clarity of gesture execution** (FC2), P3HU received the highest number of endorsements, nearly doubling the counts of PC and MV2D. Moreover, in the final instructional **ranking** (FC3), P3HU dominated the first-rank choices, while PC was overwhelmingly relegated to third place. These patterns underline the strong user preference for embodied, immersive instruction, particularly in contexts requiring precise spatial and motion understanding.

Regarding the C-SOC items, the user's scores confirm the improved social dimension of the P3HU modality. As shown in Figure 14, for both the **C-SOC1** (preferred and clearest instructional modality) and **C-SOC2** (best overall instructional experience) items, P3HU received the highest number of *rank-1* endorsements (in case of C-SOC1, the P3HU was always preferred with respect to the other two).



(a) C-SOC1. (b) C-SOC2.

Fig. 14: Rank-based comparisons for C-SOC constructs across PC, P3HU, and MV2D conditions. Rank-x indicates the number of participants who placed a modality in x-place ($x \in \{1, 2, 3\}$); stacked bars show ranking frequency.

5.5 Qualitative Analysis

To analyze in depth the aspects outlined in our quantitative analysis, and to provide a complementary perspective, we report insights inferred from participants' open-ended (qualitative) comments.

Specifically, we adopted a structured thematic coding procedure inspired by thematic and directed qualitative content analysis [22, 7]. The initial codebook was defined deductively from the analytical structure of the study, namely: (i) the three compared modalities (PC, MV2D, P3HU), (ii) the constructs already analyzed quantitatively (perceptual clarity, social presence, usability, workload, and technology acceptance), and (iii) the modality-specific interpretations. This directed content analysis allowed us to define initial codes with prior theory and considering the analytical framework of the study [4].

The resulting codes captured positive modality-related aspects, negative modality-related aspects, and cross-modality experiential limitations. Positive elements included: P3HU free-viewpoint inspection, P3HU spatial and joint-alignment clarity, P3HU embodiment support, MV2D multi-angle visibility, and PC familiarity. Negative elements instead, amounts to: MV2D confusion and cognitive overload, PC negative preference, lack of own-body feedback in immersive modalities, VR visual issues, and VR fatigue or proprioceptive discomfort. Cross-modality experiential limitations included general design requests or constraints, such as audio-guidance requests, timer and layout improvements, mixed-reality or smart-mirror suggestions.

Each response was manually inspected and assigned one or more thematic labels, allowing multiple themes to be associated with the same participant. Counts are reported as the number of unique participants mentioning each theme, considering only those who provided at least one qualitative response (each participant contributes at most once per theme, even if the same aspect was mentioned multiple times within same response). Overall, 20 out of 30 participants provided at least one open-ended comment, while 10 participants did not provide us with

any comment; therefore, the following counts are reported over the (M=20) participants who answered.
First, comments indicate that P3HU was repeatedly appreciated (14/20) for free-viewpoint inspection (3/20), improved spatial understanding of joint alignments and range of motion (6/20), and the overall sense of realistic guidance (13/20). This directly complements the quantitative findings where P3HU achieved the most favorable scores in perceived perceptual clarity and social presence, and was frequently ranked as the clearest modality for learning and remembering the technique.
Second, MV2D was positively described (5/20) as enabling multi-angle visibility and richer visual information compared to a single view, which aligns with its intermediate positioning in the quantitative comparisons for clarity-related outcomes. However, some participants reported negative aspects (3/20) related to disorientation (3/20), difficulty deciding where to look (3/20), and cognitive overload caused by the simultaneous display of multiple panels (3/20). These comments complement the quantitative results obtained with higher perceived cognitive load with MV2D.
Third, the PC baseline was described as familiar and simple (6/20), which is consistent with quantitative trends indicating PC as the easiest to use. At the same time, some participants commonly noted that PC provides limited support for depth perception and detailed joint alignment parsing (5/20), which aligns with lower perception scores and clarity rankings compared to immersive conditions.
Finally, across both immersive modalities, several users (12/20) explicitly mentioned missing visual feedback of their own body during execution (6/20) and recommended complementary design elements, including self-avatar overlays (6/20), mirrored views (1/20), skeletal cues (1/20), and audio (4/20) timing (3/20) guidance. Moreover, it is worth mentioning that a few users (4/20) outlined some issues related to immersive VR settings, including too strong virtual illumination (1/20), VR proprioceptive fatigue (2/20), and mental effort (1/20).
All those comments complements the quantitative analysis in two ways: (i) they offer a plausible explanation for improvements of P3HU modality with respect to MV2D and PC ones; and (ii) motivate concrete future directions aimed at reducing the observed limitations, considering both user interface and multimodal stimuli, to increase even more perceived self-correction ability.

6 Discussions, Limitations and Future Works

We discuss here how the obtained results answer our RQs. Concerning **RQ1** (*"When exercise content and showing timing are held constant, how do the immersive visualization modalities compare to classical flat 2D, in terms of participants' general exercise learning?"*), the results indicate that immersive guidance, and especially the P3HU condition, more effectively supports perceived exercise understanding than the classical flat-video baseline. Across PER, CM, and FC, P3HU consistently emerged as the most preferred condition, while MV2D was

generally placed between P3HU and PC. This trend was also reflected in the inferential results: both immersive modalities significantly outperformed PC on clarity-related dimensions (PER1 and PER2, all $p < .05$). In addition, P3HU received the highest number of top-ranked votes across all CM items and was also judged as the clearest and most preferred modality in FC. This pattern was also coherent with the *TAM-PU* results, which showed that both immersive modalities were systematically and statistically significant (all $p < .05$ for all items) as more useful than the PC baseline, with P3HU again showing the most favorable overall profile, even if not significantly different from MV2D. These findings suggest that, when exercise content is held constant, richer and more spatially inspectable visualizations can better support the interpretation of posture, gesture dynamics, and movement execution, with multi-view guidance already improving comprehension over flat video and embodied 3D guidance providing a further advantage.

Moving to **RQ2** (*"How do the three visualization modalities differ in terms of perceived instructor social presence, comfort, embodiment, and affective support during XR-based exercise instruction?"*), our findings indicate that the P3HU enhances the perceived social presence and affective connection with the virtual instructor. This is supported by significantly higher ratings in multiple SP items, in particular SP1, SP2, SP3, and SP5 where P3HU outperformed both MV2D and PC (all $p < .05$) and also by the C-SOC constructs (where P3HU modality always exposes the highest ranked score). We justify this considering the spatially embedded nature of the animated avatar, which is directly linked to an increase in the sense of presence and companionship. Additionally, the P3HU condition yielded significantly lower scores in both mental and physical workload compared to MV2D (NASA1 and NASA2, $p < .01$), indicating superior comfort during interaction. Moreover, it was not perceived as significantly different from the basic PC condition, which is a particularly relevant result, as it suggests that the benefits of the embodied 3D instructor were achieved without sacrificing perceived ease and interaction comfort. Moreover, it is worth mentioning that while some participants reported symptoms such as eyestrain or dizziness in their previous VR experiences (7/30), we did not detect patterns of sickness when reviewing workload scores, usability ratings, and open-ended comments. Indeed, only a few individuals noted some experiential issues during our experimental trial (only 4 participants, as reported in Section 5.5). This is crucial, considering that VR-related discomfort could influence subjective perception and so evaluations. Even though experiential issues weren't a tracked outcome, the weight of both quantitative and qualitative results did not reveal clear patterns of impact on subjective ratings. These findings are further corroborated by SUS-Positive items, where P3HU was preferred for frequency of use (SUS-POS1), and integration (SUS-POS3). Nonetheless, certain usability challenges were noted. Indeed P3HI was penalized, along with MV2D, for the need of support to use, while MV2D was considered the most complex in the SUS-Negative subscale (SUS-NEG1 and SUS-NEG2 $p < .05$) with respect to PC. Moreover, while P3HU was preferred over MV2D for ease of interaction

(TAM-PEOU1, $p < .01$), it exhibited non-significant differences for all the other constructs, reflecting users' familiarity with standard 2D videos. These results underscore a critical trade-off between immersive representational richness and cognitive accessibility, highlighting the importance of designing XR-based systems that balance experiential depth with intuitive interaction and onboarding, especially for novice users.

Considering **RQ3** (*"Does a parametric 3D avatar improve perceived clarity and social presence compared to 2D video modalities?"*), our findings provide empirical support for the instructional potential of parametric avatar-based visualization. The P3HU modality emerged as the most effective condition across all comparative constructs. Specifically, the rank-based results for the **CM-x** items (Figure 12), covering gesture comprehension (CM1), visual information quality (CM2), boundary identification (CM3), and followability (CM4) revealed that P3HU received the highest number of *rank-1* preferences across the board, clearly outperforming both MV2D and PC. Complementarily, post-experience preference data from the **FC-x** items (Figure 13) reinforced this trend. In both the preferred modality (FC1) and perceived clarity of execution (FC2), P3HU was endorsed by the majority of participants. Moreover, the final overall instructional ranking (FC3) showed P3HU consistently placed first, while the PC modality was most frequently ranked last.

These considerations about our RQs outline that employing parametric 3D models as immersive fitness guidance increases comprehension, social, and functional user perception. Following RQ1 and RQ3, both immersive modalities improved key clarity-related dimensions over PC. This pattern is coherent with prior work suggesting that XR-based guidance can improve the intelligibility of posture and movement cues, and with studies such as [9, 10] indicating the potential advantage of 3D virtual instructors over more conventional formats. However, our results extend this in a more controlled fashion, showing that such benefits persist even when the underlying exercise stimulus is fully aligned across conditions, and revealing that a multi-view representation already improves comprehension over flat video. From RQ2 and RQ3, we also understand that P3HU enhanced the perceived social presence and affective quality of the instructor, in line with prior work on co-presence, avatars, and embodied guidance [44, 14, 10]. It is worth noticing that the results from RQ2 show how such an advantage did not lower user experience. Indeed, P3HU yielded a comparable mental and physical workload to PC and was inferior to MV2D, and was positively evaluated on several usability and acceptance dimensions, whereas MV2D was more often penalized for complexity. This indicates that representational richness, if coherently organized for the user, may provide a positive impact. In this sense, the embodied avatar appears to provide not only a more complete visualization but also a more effective instructional interface, capable of supporting how users interpret, trust, and follow exercise guidance. Finally, it is worth noticing that, beyond these users' experiential factors, *XaRNold* also contributes methodologically, since its modular and data-driven use of a parametric human model avoids the need for manually recorded and registered performers

for each exercise variation, making the framework more scalable, customizable, and reusable for future experimentation and adaptive workout generation. Our study showed that avatar-based XR guidance is not simply a more engaging alternative to video, but as a structured design choice that can improve exercise comprehension, strengthen perceived companionship, and support more flexible exercise-delivery pipelines [9, 13, 10].
Despite such positive outcomes, our work has limitations that should be mentioned:

- Our contribution is focused on perceived instructional clarity and perceived social presence under aligned stimulus conditions, and do not measure actual exercise correctness, so it cannot determine whether users' improved perceptions in immersive conditions translate into better motor performance.
- The participant pool comprised mostly young and well-educated adults, a population more likely to be familiar and comfortable with immersive technologies.
- *XaRNold* best supports users with typical motor abilities and stable balance, and lacks multimodal guidance or adaptive pacing for individuals with mobility or cognitive constraints. Moreover, we included only a prototype male and female avatars, reducing the diversity of body types and cultural identities.

Considering these limitations, in future works, we will implement objective evaluation modules leveraging pose estimation to assess execution accuracy in real time and to study retention across repeated sessions. Second, we aim to integrate LLM-based adaptive workout generation by analyzing user performance logs to personalize future exercise sessions. Third, we plan to explicitly account for XR experience in follow-up experiments (e.g., stratified recruitment and/or longitudinal designs) to quantify and control potential novelty effects and extend to a more diverse demographic (including older adults or individuals with less technological exposure), pursuing generalizable insights. Finally, we will design adaptive pose generation mechanisms to dynamically adjust avatar demonstrations based on individual user needs (including inclusive support for users with physical impairments). These enhancements will enable *XaRNold* to evolve from a perceptually optimized system into an adaptive, intelligent, and inclusive training companion.

Declarations

Funding: This research received no external funding.

References

1. Abayasiri, R.A.M., Bo, A.P.L., Dick, T.J.M., Baghaei, N.: Influence of virtual-reality illusions on balance performance and immersive user experience in young adults: A within-subject experimental study. JMIR Serious Games **13**, e70376 (2025). https://doi.org/10.2196/70376

2. Ali, S.F., Azmat, S.A., Noor, A.U., Siddiqui, H., Noor, S.: Virtual reality as a tool for physical training. In: 2017 First International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT). pp. 1–6. IEEE (2017)
3. Armandi, V., Stacchio, L., Cascarano, P., Hajahmadi, S., Donatiello, L., Marfia, G.: An augmented outdoor workout system for jogging and calisthenics support. Frontiers in Virtual Reality **6**, 1613717 (2025)
4. Assarroudi, A., Heshmati Nabavi, F., Armat, M.R., Ebadi, A., Vaismoradi, M.: Directed qualitative content analysis: the description and elaboration of its underpinning methods and data analysis process. Journal of research in nursing **23**(1), 42–55 (2018)
5. Bebko, A.O., Thaler, A., Troje, N.F.: bmlsup–a smpl unity player. In: 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 573–574. IEEE (2021)
6. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
7. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative research in psychology **3**(2), 77–101 (2006)
8. Brooke, J.: Sus: A "quick and dirty" usability scale. Usability Evaluation in Industry pp. 189–194 (1996)
9. Burns, J., Xu, W., Williams, I., Khawaja, I.: Comparative study of ar versus image and video for exercise learning (2022), https://arxiv.org/abs/2209.02161
10. Chittaro, L.: Yoga in augmented reality: Comparison of a 3d experience versus traditional video of a lesson. Virtual Reality **29**, 153 (2025). https://doi.org/10.1007/s10055-025-01231-z
11. Christaki, K., Christakis, E., Drakoulis, P., Doumanoglou, A., Zioulis, N., Zarpalas, D., Daras, P.: Subjective visual quality assessment of immersive 3d media compressed by open-source static 3d mesh codecs. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.H., Vrochidis, S. (eds.) MultiMedia Modeling. pp. 80–91. Springer International Publishing, Cham (2019)
12. Clemente, F.M., Ramirez-Campillo, R., Moran, J., Zmijewski, P., Silva, R.M., Randers, M.B.: Impact of lower-volume training on physical fitness adaptations in team sports players: A systematic review and meta-analysis. Sports Medicine – Open **11**(1), 3 (2025). https://doi.org/10.1186/s40798-024-00808-3
13. Colombo, V., Mondellini, M., Aliverti, A., Sacco, M.: Immersion and interaction during cycling in virtual reality: the influence on perceived effort and subjective experience. Frontiers in Virtual Reality **Volume 6 - 2025** (2025). https://doi.org/10.3389/frvir.2025.1490588, https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2025.1490588
14. Czub, M., Janeta, P.: Exercise in virtual reality with a muscular avatar influences performance on a weightlifting exercise. Cyberpsychology: Journal of Psychosocial Research on Cyberspace **15**(3), Article 10 (Aug 2021). https://doi.org/10.5817/CP2021-3-10, https://cyberpsychology.eu/article/view/13404
15. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly **13**(3), 319–340 (1989). https://doi.org/10.2307/249008
16. Do, T.D., Protko, C.I., McMahan, R.P.: Stepping into the right shoes: The effects of user-matched avatar ethnicity and gender on sense of embodiment in virtual

reality. IEEE Transactions on Visualization and Computer Graphics **30**(5), 2434–2443 (2024). https://doi.org/10.1109/TVCG.2024.3372067
17. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)
18. Hajahmadi, S., Stacchio, L., Giacché, A., Cascarano, P., Marfia, G.: Investigating extended reality-powered digital twins for sequential instruction learning: the case of the rubik's cube. In: 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 259–268 (2024). https://doi.org/10.1109/ISMAR62088.2024.00040
19. Hamada, T., Hautasaari, A., Kitazaki, M., Koshizuka, N.: Solitary jogging with a virtual runner using smartglasses. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 644–654 (2022). https://doi.org/10.1109/VR51125.2022.00085
20. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. Human Mental Workload **1**(3), 139–183 (1988)
21. Hibbs, A., Tempest, G., Hettinga, F., Barry, G.: Impact of virtual reality immersion on exercise performance and perceptions in young, middle-aged and older adults. Plos one **19**(10), e0307683 (2024)
22. Hsieh, H.F., Shannon, S.E.: Three approaches to qualitative content analysis. Qualitative health research **15**(9), 1277–1288 (2005)
23. Huang, N., Goswami, P., Sundstedt, V., Hu, Y., Cheddad, A.: Personalized smart immersive xr environments: a systematic literature review. The Visual Computer pp. 1–34 (2025)
24. Jerald, J.: The VR book: Human-centered design for virtual reality. Morgan & Claypool (2015)
25. Jiménez-Alfageme, R., Garrone, F.P., Rodriguez-Sanchez, N., Romero-García, D., Sospedra, I., Giménez-Monzó, D., Ayala-Guzmán, C.I., Martínez-Sanz, J.M.: Nutritional intake and timing of marathon runners: Influence of athlete's characteristics and fueling practices on finishing time. Sports Medicine – Open **11**, 26 (2025). https://doi.org/10.1186/s40798-024-00801-w
26. Judd, C.M., Westfall, J., Kenny, D.A.: Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. Journal of personality and social psychology **103**(1), 54 (2012)
27. Jung, M., Sim, S., Kim, J., Kim, K.: Impact of personalized avatars and motion synchrony on embodiment and users' subjective experience: empirical study. JMIR Serious Games **10**(4), e40119 (2022)
28. Karaosmanoglu, S., Cmentowski, S., Nacke, L.E., Steinicke, F.: Born to run, programmed to play: Mapping the extended reality exergames landscape. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–28 (2024)
29. Kerby, D.S.: The simple difference formula: An approach to teaching nonparametric correlation. Comprehensive Psychology **3**(1), 11–IT (2014). https://doi.org/10.2466/11.IT.3.1
30. Kim, G., Biocca, F.: Immersion in virtual reality can increase exercise motivation and physical performance. In: International conference on virtual, augmented and mixed reality. pp. 94–102. Springer (2018)
31. Kim, H.E., Parvin, D.E., Ivry, R.B.: The influence of task outcome on implicit motor learning. eLife **8**, e39882 (apr 2019). https://doi.org/10.7554/eLife.39882, https://doi.org/10.7554/eLife.39882

32. Kittel, A., Lindsay, R., Le Noury, P., Wilkins, L.: The use of extended-reality technologies in sport perceptual-cognitive skill research: A systematic scoping review. Sports Medicine – Open **10**(1), 128 (2024). https://doi.org/10.1186/s40798-024-00794-6
33. Korkut, E.H., Surer, E.: Visualization in virtual reality: a systematic review. Virtual Reality **27**(2), 1447–1480 (2023)
34. Kourtesis, P.: A comprehensive review of multimodal xr applications, risks, and ethical challenges in the metaverse. Multimodal Technologies and Interaction **8**(11), 98 (2024)
35. Krauß, V., Boden, A., Oppermann, L., Reiners, R.: Current practices, challenges, and design implications for collaborative ar/vr application development. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2021)
36. Krein, K., Ilundáin-Agurruza, J.: High-level enactive and embodied cognition in expert sport performance. In: Sport, Ethics, and Neurophilosophy, pp. 112–126. Routledge (2020)
37. Le Noury, P., Polman, R., Maloney, M., Gorman, A.: A narrative review of the current state of extended reality technology and how it can be utilised in sport. Sports Medicine **52**(7), 1473–1489 (2022)
38. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
39. Mocco, A., Valmaggia, L., Bernardi, L., Alfieri, M., Tarricone, I.: Enhancing physical activity with immersive virtual reality: A systematic review. Cyberpsychology, Behavior, and Social Networking **27**(5), 303–317 (2024)
40. Mologne, M., Hu, J., Carrillo, E., Gomez Ladron de Guevara, D., Yamamoto, T., Lu, S., Browne, J., Dolezal, B.: The efficacy of an immersive virtual reality exergame incorporating an adaptive cable resistance system on fitness and cardiometabolic measures: A 12-week randomized controlled trial. International Journal of Environmental Research and Public Health **20**, 210 (12 2022). https://doi.org/10.3390/ijerph20010210
41. Murray, E.G., Neumann, D.L., Moffitt, R.L., Thomas, P.R.: The effects of the presence of others during a rowing exercise in a virtual reality environment. Psychology of Sport and Exercise **22**, 328–336 (2016). https://doi.org/https://doi.org/10.1016/j.psychsport.2015.09.007, https://www.sciencedirect.com/science/article/pii/S1469029215300145
42. Nehme, Y., Farrugia, J.P., Dupont, F., LeCallet, P., Lavoué, G.: Comparison of subjective methods, with and without explicit reference, for quality assessment of 3d graphics. pp. 1–9 (09 2019). https://doi.org/10.1145/3343036.3352493
43. Neumann, D.L., Moffitt, R.L., Thomas, P.R., Loveday, K., Watling, D.P., Lombard, C.L., Antonova, S., Tremeer, M.A.: A systematic review of the application of interactive virtual reality to sport. Virtual Reality **22**(3), 183–198 (2018)
44. Oh, C.S., Bailenson, J.N., Welch, G.F.: A systematic review of social presence: Definition, antecedents, and implications. Frontiers in Robotics and AI **Volume 5 - 2018** (2018). https://doi.org/10.3389/frobt.2018.00114, https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2018.00114
45. Park, S.H., Casamento-Moran, A., Singer, M.L., Ernster, A.E., Yacoubi, B., Humbert, I.A., Christou, E.A.: Integration of visual feedback and motor learning: Corticospinal vs. corticobulbar pathway. Human Movement Science **58**, 88–96 (2018). https://doi.org/https://doi.org/10.1016/j.humov.2018.01.002, https://www.sciencedirect.com/science/article/pii/S0167945717307029

46. Peña, E.A., Habiger, J.D., Wu, W.: Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. Annals of statistics **39**(1), 556 (2011)
47. Radianti, J., Majchrzak, T.A., Fromm, J., Wohlgenannt, I.: A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Computers & education **147**, 103778 (2020)
48. Razali, N.M., Wah, Y.B., et al.: Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics **2**(1), 21–33 (2011)
49. Rosner, B., Glynn, R.J., Lee, M.L.T.: The wilcoxon signed rank test for paired comparisons of clustered data. Biometrics **62**(1), 185–192 (2006)
50. Schuermans, J., Van Hootegem, A., Van den Bossche, M., Van Gendt, M., Witvrouw, E., Wezenbeek, E.: Extended reality in musculoskeletal rehabilitation and injury prevention-a systematic review. Physical therapy in sport **55**, 229–240 (2022)
51. Schulze, S., Pence, T., Irvine, N., Guinn, C.: The effects of embodiment in virtual reality on implicit gender bias. In: International Conference on Human-Computer Interaction. pp. 361–374. Springer (2019)
52. Semeraro, A., Turmo Vidal, L.: 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. pp. 1–16 (2022)
53. Stacchio, L., Balloni, E., Frontoni, E., Paolanti, M., Zingaretti, P., Pierdicca, R.: Minevra: Exploring the role of generative ai-driven content development in xr environments through a context-aware approach. IEEE Transactions on Visualization and Computer Graphics **31**(5), 3602–3612 (2025). https://doi.org/10.1109/TVCG.2025.3549160
54. Suk, H., Laine, T.H.: Influence of avatar facial appearance on users' perceived embodiment and presence in immersive virtual reality. Electronics **12**(3), 583 (2023)
55. Tavakol, M., Dennick, R.: Making sense of cronbach's alpha. International journal of medical education **2**, 53 (2011)
56. Teixeira, P.J., Carraça, E.V., Markland, D., Silva, M.N., Ryan, R.M.: Exercise, physical activity, and self-determination theory: a systematic review. International journal of behavioral nutrition and physical activity **9**(1), 78 (2012)
57. Waltemate, T., Gall, D., Roth, D., Botsch, M., Latoschik, M.E.: The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. IEEE transactions on visualization and computer graphics **24**(4), 1643–1652 (2018)
58. Wang, J., Qin, Y., Wu, Q., Zeng, D., Gao, X., Wang, Q., Li, Z., Ni, Y., Li, H., Zhang, P., et al.: An adaptive ai-based virtual reality sports system for adolescents with excess body weight: a randomized controlled trial. Nature Medicine pp. 1–14 (2025)
59. Witte, K., Bürger, D., Pastel, S.: Sports training in virtual reality with a focus on visual perception: a systematic review. Frontiers in Sports and Active Living **7**, 1530948 (2025)
60. Xiao, P.W., Fan, K.K., Xu, S., Su, C.H.: A preliminary study on the learning satisfaction and effectiveness of vr weight training assisting learning system for beginners. Eurasia Journal of Mathematics, Science and Technology Education **13**(9), 6231–6248 (2017)

61. Yannakakis, G.N., Martínez, H.P.: Ratings are overrated! Frontiers in ICT **Volume 2 - 2015** (2015). https://doi.org/10.3389/fict.2015.00013, https://www.frontiersin.org/journals/ict/articles/10.3389/fict.2015.00013