Evaluating therapist representation techniques in mixed reality-based tele-rehabilitation exergames

Evaluating Therapist Representation Techniques in Mixed Reality-based Tele-rehabilitation Exergames

Roberta Macaluso*
Politecnico di Torino, Italy

Alessandro Visconti †
Politecnico di Torino, Italy

Davide Calandra ‡
Politecnico di Torino, Italy

Roberto Ciardo §
Politecnico di Torino, Italy

Giacinto Barresi ¶
Bristol Robotics Laboratory,
UWE Bristol, UK

Fabrizio Lamberti ‖
Politecnico di Torino, Italy

Figure 1: From left to right: Audio-only, Video, and 3D Avatar representations of the remote therapist being compared.

Abstract

Recent advancements in technology have improved rehabilitation through tele-rehabilitation, offering flexible and personalized care with remote monitoring. Interactive "exergames" using eXtended Reality (XR) enhance treatment by combining traditional methods with digital gaming. However, there is limited research on virtual therapist representations. This study evaluates three techniques for therapist representation in Mixed Reality (MR) tele-rehabilitation: audio-only, video, and 3D avatar. Using a collaborative exergame for upper limb rehabilitation in Multiple Sclerosis (MS) patients, the study assesses these methods based on peer acceptance, user experience, social and co-presence, and naturalness, aiming to optimize therapist representation in MR-enabled tele-rehabilitation contexts.

Index Terms: Telerehabilitation, Extended Reality, Mixed Reality, Avatars, Exergames, Multiple Sclerosis

1 Introduction

In rehabilitation, therapist-patient relationships go beyond education and physical interventions, with holistic techniques enhancing care and communication [13]. When on-site interaction isn't possible, tele-rehabilitation offers flexible, individualized care with continuous remote monitoring thanks to recent technological advancements. eXtended Reality (XR), particularly effective in this context, creates immersive Virtual Environments (VEs) where patients can interact with a virtual therapist [4]. Specifically, Mixed Reality (MR) allows patients to engage with virtual elements and people while remaining in familiar settings like their homes.

In tele-rehabilitation, collaboration is crucial, as human-to-human interaction conveys specific sensations that facilitate effective collaborative tasks. Several studies have focused on patient interactions with virtual therapists in XR-based rehabilitation [4].

Traditionally, telerehabilitation utilized phone calls (audio) and video conferencing (audio and video) [11], allowing therapists to interact with patients and observe task execution to provide feedback. In XR-based applications, therapists are often represented as avatars [4]. These avatars, which guide patients, typically use predefined animations rather than real-time movements [7, 21, 22].

Exergames, or physical movement-based video games, widely used in XR contexts, have been shown to improve user enjoyment and reduce anxiety and stress in rehabilitation [5, 14, 17]. However, they often overlook the importance of therapist presence and guidance, which is essential for enhancing human-to-human interaction and movement assistance. Considering this, empowering therapists to personalize rehabilitation sessions with movements and feedback tailored to each patient would be valuable [4].

Building on the above considerations, this study aims to identify the most effective therapist representation in tele-rehabilitation by comparing three techniques: audio-only, video, and 3D avatar. We will evaluate aspects such as realism, attractiveness, behavior naturalness, co-presence, perceived security, engagement, and emotional/physical support. This goal is pursued using a case study of an MR-based exergame for upper limb tele-rehabilitation in Multiple Sclerosis (MS) patients. The therapist can provide real-time feedback, allowing them to tailor sessions to the patient's abilities and progress, making adjustments in difficulty as needed [8]. The hypothesis is that a 3D avatar will enhance patient engagement and co-presence more effectively than other types of representations, despite the higher fidelity of the video feed. Furthermore, it is also hypothesized that 3D avatars are more effective than audio-only or video representations in providing feedback and guidance for motor rehabilitation thanks to the possibility to display movements in three dimensions.

*e-mail: roberta.macaluso@polito.it
†e-mail: alessandro.visconti@polito.it
‡e-mail: davide.calandra@polito.it
§e-mail: roberto.ciardo@studenti.polito.it
¶e-mail: giacinto.barresi@uwe.ac.uk
‖e-mail: fabrizio.lamberti@polito.it

2 Background

A large number of studies have explored the role of XR in (tele)rehabilitation and exergames [1, 6, 14].

A notable early example is the work by Piron et al. [18], which focuses on tele-rehabilitation for post-stroke patients. They use a 3D motion tracking system to create a VE that represents the patient's movement. Simultaneously, the therapist is displayed within a Virtual Reality (VR) environment via an audio and video feed, effectively integrating traditional video conferencing into the VR experience to support the patient during exercises. Although results indicated that this method is comparable to on-site VR rehabilitation, the absence of a three-dimensional representation of the therapist may limit the effectiveness of demonstrating physical exercises involving any type of movement.

Sousa et al. [24] explore the use of Augmented Reality (AR) technology to enhance rehabilitation exercises with automatic real-time feedback. This system addresses the limitations of in-person sessions by offering an alternative to traditional tele-rehabilitation through visual guidance and automatic feedback, improving patient performance and ensuring correct movement execution while reducing injury risks. Leveraging MediaPipe to track the full-body pose of the patient, the system enables precise evaluation of exercise correctness and allows patients to identify areas for improvement. However, while automatic feedback is beneficial compared to no feedback, algorithms may struggle to accurately assess patients with varying characteristics. In such cases, having a human therapist, even remotely, provides more precise and tailored feedback for individual patients.

In this direction, Sobota et al. [23] evaluated the possibility for a therapist to connect remotely to a VR-based exergame for upper limb rehabilitation, allowing them to observe the patient and provide real-time feedback. In this setup, both users are represented by humanoid 3D avatars, with communication occurring through body language or short text messages displayed on a VR panel. The main limitations of this solution are the lack of a verbal communication channel and the inability to convey facial expressions, which are crucial for effective interaction between patient and therapist.

Although the use of virtual and remote therapists has been extensively studied, many XR exergames do not incorporate either option, focusing instead on tasks performed independently and providing feedback solely through game performance metrics. For example, Macaluso et al. [15] studied the use of wearable MR for a rehabilitation task related to multiple sclerosis. Users equipped with a Microsoft HoloLens 2 play a game designed to stimulate specific arm movements by grabbing a virtual object (a cube) and placing it onto various cells of a virtual chessboard. During the task, target cells change color over time to indicate their status, and a score is provided based on the accuracy of task execution, enhancing the gamification aspect. Later, Tanda et al. [25] explored incorporating bi-manual tasks, which are commonly required for activities of daily living, into multiple sclerosis rehabilitation. They modified the original task from Macaluso et al. [15] to require manipulation of a physical prop instead of a virtual cube. This setup used an external camera system and 2D markers to track the physical object, in addition to the HoloLens 2. In both cases, a key limitation is that the systems cannot automatically assess the quality of the movements performed by the user during rehabilitation tasks. In this context, a remote therapist could offer real-time, precise feedback and correct execution errors, thereby enhancing the tool's effectiveness.

There are numerous ways to represent avatars in multi-user XR contexts [26, 27] and various methodologies for depicting virtual or remote therapists in tele-rehabilitation [4]. Some studies have explored non-human representations to make the experience more game-like [17], but such representations are often deemed unsuitable for rehabilitation tasks [4]. For example, in [20], poor avatar characterization (represented as spheres) led to user discomfort. Therefore, identifying the most suitable VR avatar representation for a specific use case is not straightforward [3].

Based on this review, the proposed work compares two methodologies for representing remote therapists in XR-based tele-rehabilitation: video conferencing and 3D avatars. To provide a comprehensive evaluation, a third audio-only variant, a traditional method in non-VR tele-rehabilitation [11], is also included. The study uses the exergame from [15, 25], as it represents typical rehabilitation tasks for the medical condition and is suitable for incorporating remote supervision by a therapist.

3 Materials and Methods

This section introduces the exergame that has been developed for the study and provides implementation details for the considered therapist representations.

3.1 Original Exergame Design

As mentioned in Section 2, the exergame considered in this study draws inspiration from the one investigated in [15, 25], which is based on the repeated placement of a virtual object (i.e., a yellow cube) on a virtual chessboard (organized as a 3×3 grid). The player (patient), equipped with a Microsoft HoloLens 2, can interact with the cube and the grid tiles using hand tracking. Initially, a random tile becomes active and starts glowing green. An active tile represents a correct target for the rehabilitation movement performed with the cube. The patient scores 1 point for each green tile the cube moves over. Other tiles, glowing red, serve as incorrect targets and trigger audio-visual feedback if the cube reaches them, although these errors do not result in a point deduction. Active tiles alternate between green and red every two seconds, with a purple indicator during the transition. The objective of the game is to achieve the highest score by guiding the cube across the green tiles. This set of movements also constitutes the rehabilitation routine of the exergame.

3.2 Modifications to the Exergame

The new iteration of the exergame was developed, using Unity 2022.3 and MRTK[1] v2.7, similar to the original work [15].

For the purpose of the study, the original application has been integrated with multi-user capabilities. To this aim, Mirror[2], a high-level networking library for Unity was used to create a client-server architecture letting users (i.e., patients, playing the game, and remote therapists) to connect to a shared VE. The objective is to enable the therapist to observe the patient during the experience, as well as to provide real-time feedback and support the correct execution. The Mirror server is hosted on the desktop PC, which runs the therapist's side of the multiplayer application. The patient, identified as a client, joins the experience through an instance deployed and running on an MR device.

Regardless of the therapist's representation, the following key features were be implemented to enable therapist-patient interaction:

- *Visualization of the patient:* the therapist, from his or her point of view, can see only a simple representation of the patient in which head and hands are visualized through cube objects moving in real time (Fig. 2). Patient's hand and head positions and rotations were managed using Mirror's SyncVars, which allow automatic synchronization between the server side and client side. As soon as the patient moves, the server updates the SyncVars, propagating the updates to the client(s).

[1]MRTK: https://tinyurl.com/43xhv4tf
[2]Mirror: https://mirror-networking.com/

- *Audio transmission:* voice transmission was handled using the Dissonance VOIP plugin[3], a high-quality, low-latency voice chat system for Unity, ideal for MR applications. This choice was motivated by the high flexibility and ease of integration with Mirror.

3.3 Therapist Representation Techniques

In the following, details for three considered representation techniques for the remote therapist are discussed (Fig. 1):

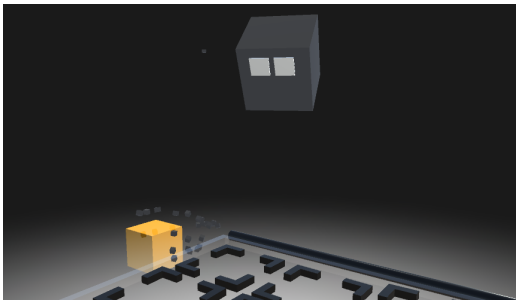


Figure 2: Representation of the patient's head and hands as seen by the remote therapist.

- The *Audio-only* technique is the simplest from an implementation perspective. It requires only the addition of a bidirectional audio channel to the original exergame's functionalities, a feature common to all three conditions. As mentioned earlier, this mode was included because it is considered a traditional method for tele-rehabilitation [11].

- Video conferencing, hereafter referred to as *Video*, is one of the most common configurations in both traditional tele-rehabilitation [11] and VR contexts [18]. It is achieved by synchronizing a video feed with the audio stream. Originally designed for desktop screens, in VR it involves simulating a 2D screen within the environment to display the video feed. Unlike the audio-only setup, this allows the patient to see the therapist in real-time through a webcam. The MixedReality WebRTC[4] library was used for video and audio streaming, avoiding Dissonance for the audio transmission to prevent desynchronization.

- The last technique, labeled *3D Avatar*, allows for the visualization of a therapist's realistic representation in the virtual scene. In literature, various methods for controlling an avatar representation have been proposed, ranging from predefined playback of animations to real-time tracking using motion capture systems. Additionally, avatars can exhibit different levels of expressiveness, from featureless faces to facial movement approximation algorithms, and even real-time facial tracking. In general, full-body realistic avatars [3] with direct facial tracking [26] are perceived as superior from various perspectives. For this reason, it was decided to opt for a realistic looking full-body representation, with the support of facial tracking.

To achieve these functionalities required for the 3D Avatar technique, while minimizing the need for expensive hardware and complex setup for the therapist, the Google MediaPipe computer vision suite was chosen. Using MediaPipe, a standard webcam can provide full-body skeleton tracking of the therapist, which can then be applied to any appropriately rigged virtual avatar. It also allows for capturing a texture of the therapist's face from a photo to apply to the avatar (by means of MediaPipe Face Mesh[5]) and enables real-time tracking of facial expressions.

The MediaPipe Unity Plugin[6], was employed to integrates MediaPipe framework with Unity. MediaPipe provides specialized machine learning-based components for tracking and detecting various body parts using landmark-based approaches by the use of a simple webcam. The following components were utilized: *Pose Detection*, which tracks 33 body landmarks for applications in fitness and rehabilitation, *Hand Detection*, which monitors 21 landmarks per hand for gesture control and sign language recognition, *Face Detection*, which identifies 468 facial landmarks for real-time expression analysis, and *Blendshape Detection*, which captures predefined facial expressions for animating digital avatars. Although the therapist is expected to use the application while seated in front of the PC, the pose detection and hand detection components experienced difficulties in estimating depth due to the limitations of the single-camera setup.

After preliminary tests with the setup, it was decided to integrate an additional specialized device for upper limb tracking, specifically the Leap Motion Controller. Placed on the therapists desk, this device provided superior tracking of the hands and upper limbs. When available, its data were used to override the less accurate tracking from MediaPipe. This adjustment was essential to achieve the performance needed for the therapist to demonstrate the required movements effectively.

4 Experiments

This section describe the study that was conducted on healthy subjects to gather a preliminary feedback on the therapist representation techniques.

4.1 Subjects

Fifteen volunteers (11 males and 4 females, aged 25.80 ± 1.90) were recruited from among students and staff at Politecnico di Torino. Since the study aims to evaluate the therapist's representation rather than the effectiveness of the rehabilitation task or the patient's representation, it was not necessary to involve an actual therapist or patients with MS. Consequently, the tests were conducted with healthy participants and an experimenter serving as the therapist. Participants were asked to sign an informed consent form based on the Helsinki Declaration too.

4.2 Procedure

The experiments followed a within-subjects design with three conditions corresponding to the three therapist representation techniques under consideration. These conditions were administered consecutively to all subjects, with the presentation order following a Latin square design to mitigate possible learning effect. Each condition included the following phases:

1. *Calibration Phase* (CP): the subject performs the calibration of the HoloLens 2 to accurately track eye movements, ensure correct hand detection, and properly align the holograms.

2. *Explanation Phase* (EP): the experimenter playing the role of the therapist provides a detailed explanation of the game to the subject, introducing the basic rules and the specific movements he or she will be required to mimic while crossing the green tiles with the cube. These explanations should contribute to making the game self-explanatory and easy to use as recommended in [19].

[3]Dissonance Voice Chat: https://tinyurl.com/mr3wk4zj
[4]MixedReality WebRTC: https://tinyurl.com/yn2d8437
[5]MediaPipe Face Mesh: https://tinyurl.com/524wxe9y
[6]MediaPipe Unity Plugin: https://tinyurl.com/4n5x22eh

3. *Gameplay Phase* (GP): the subject plays the game, focusing on executing the correct movements. During this phase, the experimenter playing the role of the therapist provides positive or negative feedback regarding the accuracy of the movements. This approach facilitated the simulation of a patient-therapist interaction scenario.

The choice of the three conditions was motivated by the intention to compare the possible advantages of a 3D avatar representation realistically reproducing the remote therapist (both in terms of appearance and movements) against common distant communication means, i.e., audio and video. The therapist-patient interaction approaches used in each of the three configurations are the following:

- *Audio-only*: during the EP, the therapist gives verbal instructions on game mechanics, specifying objects to interact with and movements to perform. During GP, feedback is provided only through these verbal cues.
- *Video*: during the EP, the subject watches a video stream of the therapist who provides the same instructions delivered in the audio-only condition. In the GP, the therapist provides feedback through both verbal and 2D visual instructions.
- *3D Avatar*: during the EP, the subject sees a real-time animated 3D avatar of the therapist providing instructions, indicating objects, and demonstrating movements, utilizing the virtual space's depth cues. In the GP, the therapist provides feedback to the subject through verbal instructions, as well as via 3D visual and virtually co-located 3D instructions.

4.3 Metrics

Both subjective and objective metrics were used to gather a comprehensive picture of the subjects' interaction experience and performance.

4.3.1 Subjective Metrics

Subjective metrics were collected through a set of questionnaires, which were administered at the end of the EP or GP (or both), that are reported below: The evaluation of the therapist's representation involved several scales. The *Godspeed Scale* [2] was utilized to assess the naturalness, acceptability, movement quality, expressiveness, and overall attractiveness of the therapists appearance, focusing specifically on the EP. For the GP, the *AttrakDiff* [10] was employed to gauge the perceived user experience, concentrating on usability, aesthetics, and overall appeal.

To evaluate the level of co-presence experienced during interactions with the remote therapist, the *Networked Minds Social Presence Questionnaire* (NMMSQ) [9] was used. Additionally, the *Behavior Naturalness* [12] scale assessed how natural the verbal and non-verbal behaviors of the therapists representation appeared to the subjects. The *Social Presence* [16] scale measured the subjects' perception of the therapists presence and their feelings of engagement and support.

Finally, a set of *Custom Questions* was included to address specific aspects of the experience, such as the clarity of the therapists explanations, assistance in understanding the required movements, potential distractions caused by the therapists representation, attention maintenance, and the appropriateness of the representation in relation to the explanations provided.

4.3.2 Objective Metrics

Objective metrics were gathered by observing subjects' behavior during interactions with the game and the remote therapist. Eye-tracking data was used to quantify visual attention, focusing on two main metrics: *timeGrid*, the duration subjects spent looking at the game grid, and *timeAvatar*, the duration spent observing the therapist's representation (if present).

5 Results

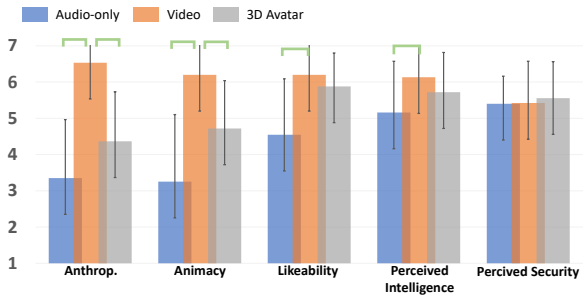
In this section, subjective and objective results are presented.

5.1 Subjective Results

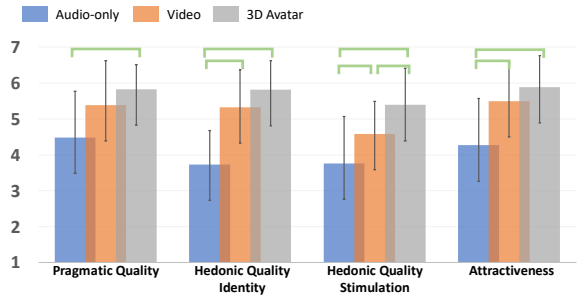
To analyze subjective metrics, data normality was assessed using the D'Agostino-Pearson test. For normally distributed data, ANOVA determined significance, with subsequent T-tests examining differences between the conditions (audio-only, video, and 3D avatar). For data not following a normal distribution, the Friedman test was used for significance, and pairwise comparisons were conducted using the Wilcoxon test. Bonferroni correction was applied to both Wilcoxon and T-tests with a significance level of $p < .016$.

In the evaluation phase (EP), the Godspeed Scale (Fig. 3a) revealed a notable preference for the video condition over the audio-only and 3D avatar conditions in terms of anthropomorphism, animation, likability, and perceived intelligence. The audio-only condition was generally regarded as the least realistic. However, no significant differences were observed regarding perceived security among the three conditions.

In the general phase (GP), analysis using the AttrakDiff (Fig. 3b) showed a significant preference for the 3D avatar over both the audio-only and video conditions in terms of stimulation (3D avatar vs. video: $p = .006$, 3D avatar vs. audio-only: $p < .001$). The 3D avatar was rated as more inventive, creative, innovative, and novel, than the other two. No significant differences were found between the video and 3D avatar conditions regarding identity, indicating that subjects felt equally able to assert their identity in both setups. While no significant differences in pragmatic quality were observed between the audio-only and video conditions, the 3D avatar significantly outperformed the audio-only condition ($p < .001$), suggesting that the 3D avatar effectively overcomes the limitations of the audio-only setup, thereby providing better support for the task.



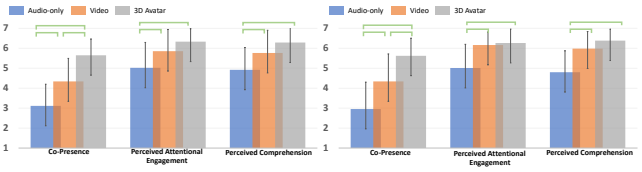
(a) Godspeed



(b) AttrakDiff

Figure 3: Results for the (a) Godspeed Scale (EP), and (b) the AttrakDiff (GP). Brackets denote statistically significant differences ($p < .016$), whereas bars represent standard deviations.

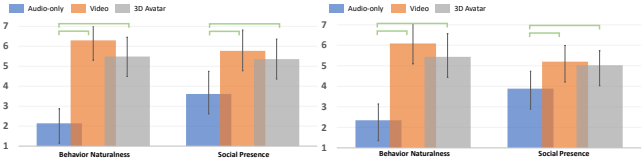
In both phases, the NMMSQ results (Fig. 4) showed a significant increase in the sense of co-presence with the 3D avatar compared to the video condition (EP: $p = 0.001$, GP: $p < .001$). For perceived attentional engagement, both the video and 3D avatar conditions were rated better than the audio-only condition in the EP (video vs audio-only: $p < .001$, 3D avatar vs audio-only: $p = .004$) and the GP (video vs audio-only: $p = .001$, 3D avatar vs audio-only: $p < .001$). Similarly, perceived comprehension was also better for the 3D avatar and video conditions compared to audio-only during both the EP (video vs audio-only: $p < .001$, 3D avatar vs audio-only: $p < .001$) and GP (video vs audio-only: $p = .003$, 3D avatar vs audio-only: $p < .001$). No significant differences were found between the video and 3D avatar conditions in terms of comprehension, aligning with the AttrakDiff findings regarding identity.



(a) EP (b) GP

Figure 4: NMMSQ results: (a) EP, and (b) GP. Brackets denote statistically significant differences ($p < .016$), whereas bars represent standard deviations.

For Behavior Naturalness (Fig.5), the results from both phases revealed that the video and 3D avatar conditions scored significantly higher than the audio-only condition (video vs audio-only: $p < .001$, 3D avatar vs audio-only: $p < .001$). No significant differences were observed between the video and 3D avatar conditions. Similarly, in terms of Social Presence (Fig.5), there were no significant differences between the video and 3D avatar conditions. However, both conditions showed significantly higher scores compared to the audio-only condition in both the EP (video vs audio-only: $p < .001$, 3D avatar vs audio-only: $p < .001$) and the GP (video vs audio-only: $p < .001$, 3D avatar vs audio-only: $p < .001$).



(a) EP (b) GP

Figure 5: Results regarding Behavioral Naturalness and Social Presence for the (a) EP, and (b) GP. Brackets denote statistically significant differences ($p < .016$), whereas bars represent standard deviations.

For the custom questions (Fig. 6), notable findings emerged from question #8 of the GP (Emotional Impact). Significant differences were found between the 3D avatar condition and both the audio-only condition ($p = .007$) and the video condition ($p = .015$), with no significant difference observed between the video and audio-only conditions. For question #7 (Immersion), the 3D avatar condition resulted in significantly better values compared to the audio-only condition ($p < .001$). Similarly, for question #2 (Movement Guidance), the 3D avatar condition was significantly better than the audio-only condition ($p < .001$), with the same result observed for both phases. Notably, both audio-only and 3D avatar conditions were rated higher than the audio-only condition during the EP as well (video vs audio-only: $p < .001$, 3D avatar vs audio-only: $p < .001$).

5.2 Objective Results

The objective measures revealed distinct viewing patterns during the EP. In the audio-only condition, subjects spent only 60% of their time looking at the grid, despite the lack of a visual representation of the therapist. In contrast, in the video condition, subjects allocated 17.3% of their time to the grid and 55.2% to the video feed, totaling 72.5% of their time engaged with virtual elements. In the 3D avatar condition, subjects spent 27.1% of their time looking at the grid and 49.2% at the 3D avatar, resulting in a total of 76% of their time focused on virtual elements. Notably, the time spent looking at the grid was significantly higher in the 3D avatar condition compared to the video condition ($p < .001$).

During the GP, subjects in the audio-only condition spent 59.9% of their time looking at the grid, a figure similar to that observed in the EP. In the video condition, however, the time spent on the grid increased compared to the EP. For the 3D avatar condition, the increase in time spent looking at the grid from the EP to the GP was smaller (EP vs GP: $p = .003$) compared to the increase seen in the video condition (EP vs GP: $p < .001$). Although the time spent looking at the avatar generally decreased from the Explanation to the GP, subjects still spent more time looking at the 3D avatar (24.6%) than the video feed (21.1%).

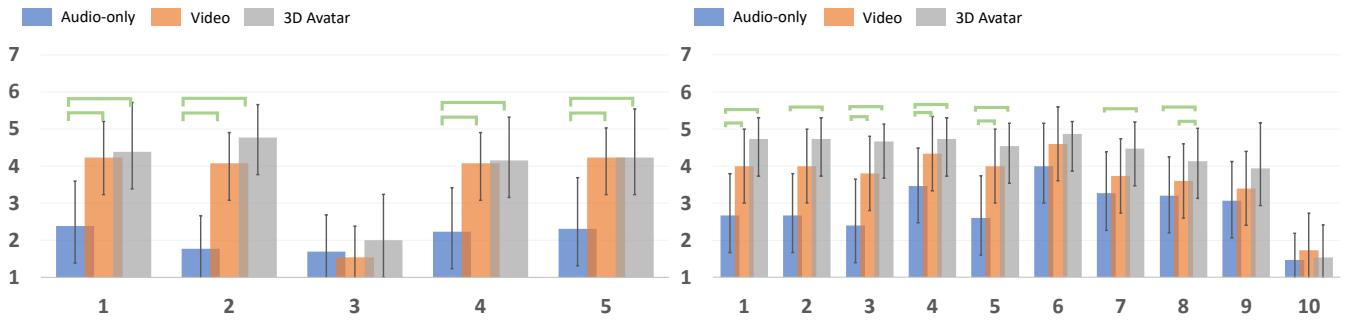
6 Discussion

The data analysis revealed that video configurations generally offered a stronger sense of realism, making participants feel as though they were interacting with a real person. This aligns with previous research indicating that video is perceived as more realistic than 3D avatars, which often struggle to consistently convey the therapists intentions, as confirmed by the Social Presence questionnaire. However, no significant differences in perceived security were found among the three conditions, suggesting that none of the representations made users feel unsafe. Despite its lower realism, the 3D avatar did not negatively impact user comfort and was comparable to having no representation at all.

The absence of a significant difference in perceived intelligence between the video and 3D avatar conditions can be attributed to the avatars inconsistent animation and communication, which led participants to perceive it as less intelligent. Nevertheless, the 3D avatar outperformed the audio-only condition in terms of pragmatic quality, unlike the video. This suggests that while the 3D avatar may be perceived as less intelligent, its immersive 3D interaction provided better task support compared to the 2D video feedback.

During the EP, users, despite being more attentive to the therapists explanation in the video configuration, spent significantly less time observing the grid compared to the 3D avatar configuration. This occurred because the video setup lacked the immersive 3D experience, forcing participants to choose between focusing on the 2D panel or the game movements. Consequently, they tended to ignore the therapist when performing movements and focused more on the therapist during explanations, which were presented on a less immersive plane. In contrast, the 3D avatar, by fully integrating into the 3D space, kept users more engaged and attentive, as confirmed by the objective metrics showing that the 3D avatar offered a more balanced and engaging experience.

Similarly, during the GP, time measurements revealed that the time spent observing the grid was similar across all configurations, with no significant increase in the audio-only condition. This indicates that users are easily distracted from the game regardless of whether a 3D avatar, video, or no therapist representation is present. This finding is further supported by question #7 (GP), where the



(a) EP

(b) GP

Figure 6: Results for the Custom Questions included in the questionnaire. EP: #1: Clarity of Explanation #2: Movements Comprehension #3: Distraction #4: Attention Maintenance #5: Representation Appropriateness GP: #1: Correction Clarity #2: Movement Guidance #3: Support #4: Mistake Correction #5: Motivation #6: Focus #7: Immersion #8: Emotional Impact #9: Time Perception #10: Concentration Difficulty Brackets denote statistically significant differences ($p < .016$), whereas bars represent standard deviations.

3D avatar significantly outperformed the audio-only condition in terms of immersion, while the video representation did not. The results suggest that the 3D avatar enhances user immersion and engagement, making them feel more involved and in control. Overall, therapist interventions during GP enhance game efficiency by providing interactive feedback and helping users maintain focus.

The Behavior Naturalness results revealed significant differences between visual and audio-only representations, but no notable differences between the video and 3D avatar representations. This indicates that, despite its lower realism, the 3D avatar's expressiveness was comparable to that of the video. The NMMSQ results further support this, showing that the 3D avatar effectively communicated its intentions better than other representations. Thus, while the 3D avatar achieved a similar level of naturalness to the video, it also enhanced co-presence, demonstrating overall superiority in representation.

Additionally, responses to the custom questions indicated that the 3D avatar significantly improved the user experience during gameplay. It positively affected mood, made the game feel less mechanical, and provided reassuring stimuli, making users feel more secure and relaxed. Notably, while any visual feed in the EP aids in understanding movements, only the immersive 3D avatar representation proved effective in engaging users during the GP. Additionally, the results from the custom question #2 (EP) suggest that while any visual feed helps with understanding movements, only an immersive 3D avatar representation proved effective in engaging users during the GP.

7 Conclusion

This study aims to identify the most effective techniques of therapist representation in XR tele-rehabilitation by evaluating aspects such as realism, co-presence, engagement, and the ability to provide feedback during rehabilitation tasks. To achieve this, an existing MR exergame for MS rehabilitation was used as test bench [15]. Originally designed to ease the psychological burden of traditional rehabilitation, the exergame was enhanced to incorporate the therapist directly into the game. This adaptation sought to boost patient engagement, offer emotional and physical support, and provide feedback to motivate persistence. Specifically, the study compared three therapist representations: audio-only, video, and 3D avatar. In each representation, real-time support was provided: vocal feedback in the audio-only case, video and audio feedback in the video case, and a real-time 3D avatar in the third case, mimicking the therapist's movements and facial expressions.

The results revealed positive feedback regarding the presence of a support figure within the game. Among the different representations, the fully immersive 3D avatar was generally preferred over the less immersive but more realistic video support, while the audio-only support was less favored. The video representation offered a higher sense of realism compared to the 3D avatar; however, the avatar did not compromise user comfort, suggesting that lower realism does not negatively impact the sense of security. Therefore, the integration of the 3D avatar into the MR environment created a more balanced and immersive experience, making users feel more involved and in control.

Despite its limitations in realism, the 3D avatar's expressiveness was comparable to that of the video and even enhanced the sense of co-presence beyond what the video could achieve. This made the 3D avatar superior in representing the therapist's presence, as it communicated intentions more effectively than the other techniques. Additionally, the 3D avatar improved the overall user experience by boosting mood, reducing the mechanical feel of the game, and providing positive stimuli that increased feelings of security and pleasure. It was especially effective during the GP, offering valuable engagement and support without distracting users from their rehabilitation tasks.

In conclusion, while the 3D avatar may offer lower or similar levels of realism compared to other methods, it excels in providing an immersive experience, increasing user engagement, and effectively supporting therapeutic tasks. This indicates that XR tele-rehabilitation systems should prioritize 3D avatars to optimize patient involvement and therapeutic outcomes. Future research should aim to enhance the realism and animation consistency of 3D avatars or investigate about creating a hybrid version that leverages the advantages of both video and 3D avatar techniques.

Acknowledgments

This work has been carried out in the frame of the VR@POLITO initiative. This investigation was also part of ENACT, a FISM (Fondazione Italiana Sclerosi Multipla) Special Project. ENACT is supported by FISM cod. 2021/Special/003 and financed with the "5 per mille" public funding, with IIT (Istituto Italiano di Tecnologia) co-funding.

REFERENCES

[1] A. Asadzadeh, T. Samad-Soltani, Z. Salahzadeh, and P. Rezaei-Hachesu. Effectiveness of virtual reality-based exercise therapy in rehabilitation: A scoping review. *Informatics in Medicine Unlocked*, 24:100562, 2021. 2

[2] C. Bartneck. Godspeed questionnaire series: Translations and usage. In *Int. Handb. of Behavioral Health Assessment*, pp. 1–35. Springer, 2023. 4

[3] D. Calandra, F. G. Pratticó, G. Lupini, and F. Lamberti. Impact ofa-vatar representation inavirtual reality-based multi-user tunnel fire simulator fortraining purposes. In *Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 1691, pp. 3–20, 2023. 2, 3

[4] S. E. Crowe, M. Yousefi, B. Shahri, T. Piumsomboon, and S. Hoermann. Interactions with virtual therapists during motor rehabilitation in immersive virtual environments: A systematic review. *Frontiers in Virtual Reality*, 5:1284696, 2024. 1, 2

[5] P. Elena, S. Demetris, M. Christina, and P. Marios. Differences between exergaming rehabilitation and conventional physiotherapy on quality of life in parkinson's disease: A systematic review and meta-analysis. *Frontiers in Neurology*, 12:683385, 2021. 1

[6] T. C. Elliott, J. D. Henry, and N. Baghaei. Designing humanoid avatars in individualised virtual reality for mental health applications. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 315–321, 2023. 2

[7] M. Fantini. *A low-cost multi-sensory wrist-mounted haptic device*. PhD thesis, Rice University, 2022. 1

[8] N. Hamzeheinejad, D. Roth, S. Monty, J. Breuer, A. Rodenberg, and M. E. Latoschik. The impact of implicit and explicit feedback on performance and experience during VR-supported motor rehabilitation. In *Proc. of 2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 382–391, 2021. 1

[9] C. Harms and F. Biocca. Internal consistency and reliability of the networked minds measure of social presence. In M. Alcaniz Raya and B. Rey Solaz, eds., *7th Ann. Int. Workshop: Presence*. UPV, 2004. 4

[10] M. Hassenzahl. Hedonic, emotional, and experiential perspectives on product quality. *Encycl. of Human Computer Interaction*, 1:266–272, 2006. 4

[11] R. S. Hinman, B. J. Lawford, and K. L. Bennell. Harnessing technology to deliver care by physical therapists for people with persistent joint pain: Telephone and video-conferencing service models. *J. of Applied Biobehavioral Research*, 24(2):e12150, 2019. 1, 2, 3

[12] P. Kullmann, T. Menzel, M. Botsch, and M. E. Latoschik. An evaluation of other-avatar facial animation methods for social VR. In *Ext. Abst. of 2023 CHI Conf. on Human Factors in Computing Systems (CHI EA '23)*, pp. 1–7, 2023. 4

[13] K. J. Loomis, S. C. Roll, and M. E. Hardison. The role of therapist-patient relationships in facilitating engagement and adherence in upper extremity rehabilitation. *Work*, 76(3):1083–1098, 2023. 1

[14] E. A. Lorenz, A. Bråten Støen, M. Lie Fridheim, and O. A. Alsos. Design recommendations for XR-based motor rehabilitation exergames at home. *Frontiers in Virtual Reality*, 5:1340072, 2024. 1, 2

[15] A. Macaluso, A. Bottino, F. G. Pratticó, F. Lamberti, C. Galletti, C. Storchi, J. Podda, A. Tacchino, G. Brichetto, N. Boccardo, L. D. Michieli, and G. Barresi. Executive control in a mixed reality exergame for motor-cognitive rehabilitation in multiple sclerosis. In *Proc. of 2024 IEEE Int. Conf. on Consumer Electronics (ICCE Las Vegas)*, pp. 1–6, 2024. 2, 6

[16] G. Makransky, L. Lilleholt, and A. Aaby. Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, 72:276–285, 2017. 4

[17] Z. Mihajlovic, S. Popovic, K. Brkic, and K. Cosic. A system for head-neck rehabilitation exercises based on serious gaming and virtual reality. *Multimedia Tools and Applications*, 77:19113–19137, 2018. 1, 2

[18] L. Piron, A. Turolla, P. Tonin, F. Piccione, L. Lain, and M. Dam. Satisfaction with care in post-stroke patients undergoing a telerehabilitation programme at home. *J. of Telemedicine and Telecare*, 14(5):257–260, 2008. 2, 3

[19] A. Schättin, S. Häfliger, A. Meyer, B. Früh, S. Böckler, Y. Hungerbühler, E. D. de Bruin, S. Frese, R. S. Egli, U. Götz, R. Bauer, and A. L. Martin-Niedecken. Design and evaluation of user-centered exergames for patients with multiple sclerosis: Multilevel usability and feasibility studies. *JMIR Serious Games*, 9(2):e22826, 2021. 3

[20] S. H. H. Shah, A. S. T. Karlsen, M. Solberg, and I. A. Hameed. A social VR-based collaborative exergame for rehabilitation: Codesign, development and user study. *Virtual Reality*, 27(4):3403–3420, 2023. 2

[21] B. Sobota, . Koreko, J. Gvuov, and M. Mattov. Therapist-patient interaction in virtual reality at the level of the upper limbs. In *Proc. 2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pp. 584–588, 2022. 1

[22] B. Sobota, . Koreko, J. Gvuov, and M. Mattov. Therapist-patient interaction in virtual reality at the level of the upper limbs. In *Proc. of 2022 20th Int. Conf. on Emerging eLearning Technologies and Applications (ICETA)*, pp. 584–588, 2022. 1

[23] B. Sobota, . Koreko, J. Gvuov, and M. Mattov. Therapist-patient interaction in virtual reality at the level of the upper limbs. In *Proc. of 2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pp. 584–588, 2022. 2

[24] M. Sousa, J. a. Vieira, D. Medeiros, A. Arsenio, and J. Jorge. SleeveAR: Augmented reality for rehabilitation using realtime feedback. In *Proc. of the 21st International Conference on Intelligent User Interfaces (IUI '16)*, pp. 175–185, 2016. 2

[25] M. Tanda, F. G. Prattic, J. Podda, E. Grange, G. Brichetto, L. De Michieli, F. Lamberti, and G. Barresi. Rehabilitative exergaming in multiple sclerosis: Bimanual tasks in mixed reality. In *Proc. of 2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, pp. 1–6, 2024. 2

[26] A. Visconti, D. Calandra, and F. Lamberti. Comparing technologies for conveying emotions through realistic avatars in virtual reality-based metaverse experiences. *Computer Animation and Virtual Worlds*, 34(3-+-4):e2188, 2023. 2, 3

[27] F. Weidner, G. Boettcher, S. A. Arboleda, C. Diao, L. Sinani, C. Kunert, C. Gerhardt, W. Broll, and A. Raake. A systematic review on the visualization of avatars and agents in AR & VR displayed using head-mounted displays. *IEEE Tran. on Visualization and Computer Graphics*, 29(5):2596–2606, 2023. 2