The Avatarm: Interacting in the Physical Metaverse via Robotics, Diminished Reality, and Haptics

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Digital Object Identifier 10.1109/ACCESS.2024.3420717

The Avatarm: Interacting in the Physical Metaverse via Robotics, Diminished Reality, and Haptics

BERNARDO BROGI [1,*], GIOVANNI CORTIGIANI [1,*], ALBERTO VILLANI[1], NICOLE D'AURIZIO[1,2], DOMENICO PRATTICHIZZO[1,2], AND TOMMASO LISINI BALDI[1,2]

[1] Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy
[2] Humanoids and Human Centered Mechatronics, Istituto Italiano di Tecnologia, 16163 Genoa, Italy

Corresponding author: Bernardo Brogi (bernardo.brogi@student.unisi.it)

This work was supported in part by the European Union's Horizon Europe Program through the Project ''HARIA—Human-Robot Sensorimotor Augmentation'' under Grant 101070292, and in part by Leonardo S.p.A. under Grant LDO/CTI/P/0025793/22.

Bernardo Brogi and Giovanni Cortigiani contributed equally to this work.

* **ABSTRACT** The Metaverse is a three-dimensional digital space where users interact and communicate through their avatars, creating a sense of presence and immersion. The use of avatars allows to convey body language and to establish an object-based dialogue, improving the expressiveness of communication between users. However, existing metaverses are limited to digital interaction, as the avatar is not able to touch, perceive, grasp or manipulate the physical objects surrounding the remote interlocutor. In a previous work, we laid the foundation for the ''Avatarm'', an avatar able to manipulate both the physical and the digital worlds. This was achieved through a robotic manipulator that remains hidden from the user's view by means of diminished reality techniques. Building upon this groundwork, in this work we have advanced the capabilities of the Avatarm with the integration of force and vibrotactile haptic feedback, empowering the user to tangibly perceive manipulated objects, and gain a heightened sense of situational awareness. In addition, to increase the realism of the interaction, we have implemented a robot control algorithm capable of matching the position and orientation of physical objects with their digital counterparts. Furthermore, a digital avatar has been integrated into the augmented environment along with a newly designed virtual hand for more realistic manipulation of virtual objects. The enhanced version of the ''Physical Metaverse'' is tested in this work through an extensive experimental campaign considering both objective and subjective measures of performance.

INDEX TERMS Haptic interfaces, hardware and software that enable touch-based interactions with real, remote, and virtual environments, manipulators, virtual reality.

I. INTRODUCTION

The term ''Metaverse'' was coined exactly 30 years ago by Neal Stephenson. In his famous novel *Snow Crash*, he described a digital urban environment where users appear through their avatars and have first-person experiences. Although Stephenson gave a negative connotation to the concept of the metaverse, describing it as a dystopian world, today the perception we have of it has significantly changed. Metaverse popularity is rapidly growing thanks to a user base

FIGURE 1. Physical Metaverse representative scenario. Two users (User A and User B) are in different locations and can interact in a shared environment of the metaverse, which is an augmented version of User A's physical surroundings (top left panel). Through a head-mounted display, User A observes his workspace (middle left panel) diminished of the robot, and augmented with the digital twin of the pitcher and the virtual avatar of User B (bottom left panel). At the same time, User B (top right panel) observes the same environment from a remote camera that streams into his head-mounted display (middle right panel). The robot is removed from the video stream and substituted with User B's virtual arm (bottom right panel). User B controls the Avatarm, hence can manipulate real objects in the physical environment of User A and feel the interaction through a wearable haptic interface.

increasingly accustomed to the 'digital lifestyle', creating a favorable environment for its commercialization as a broader application for business and life. Thanks to the metaverse, people can now interact in shared environments regardless of their geographic location, engaging in a significantly more immersive experience than that provided by conventional teleconferencing tools [1]. This is achieved through the use of Extended Reality (XR) technologies, i.e. Augmented, Mixed and Virtual Reality, thanks to which the user is superimposed or even fully immersed in a virtual environment that enriches or substitutes the physical reality of its body and its surroundings [2].

The idea of experiencing an immersive digital environment is actually older than Stephenson's definition a nd c an be traced back to early video games. However, the profoundly different philosophy behind today's metaverse is the concept of interreality [3], which emerged from the e-health field. Interreality expresses the twofold link between the physical and the digital worlds. What happens in the real world influences the experience in the virtual world and vice-versa, leading to a seamless fusion of the two environments. The pairing is enabled by the "digital twins", virtual representations of real-world entities synchronized with their physical counterparts. This deep connection between the two worlds addresses issues related to the "eternal digital present" of digital worlds [4] and contextually opens up an entirely new portfolio of applications, ranging from work to entertainment, from healthcare to education.

Despite its commercial and social potential, methodologies for physical-virtual interactivity in shared environments, particularly regarding the manipulation of objects among users, remain underexplored. Advancements in this direction could expand the possibilities of the metaverse towards a new paradigm for physical interaction in XR, which we like to refer to as the *Physical Metaverse* [5].

To comprehend the current limitation and potential of XR-mediated experiences, let us consider two friends, separated by distance, who wish to enjoy a cup of tea together. The simplest solution would involve arranging a video call and separately filling their cups, but this would likely result in the least engaging experience. Now, imagine both friends wearing head-mounted displays (HMDs) that allow them to see their living room augmented with their friend's avatar. In this scenario, each friend is physically seated at their table, but virtually positioned in front of their friend. This aligns with Mark Zuckerberg's vision for the metaverse in the coming years, as he expressed in his video keynote published in October 2021.[1] However, this implementation of the metaverse does not allow the two friends to manipulate objects on each other's tables simultaneously, as they cannot cross the boundaries of the digital space to reach and move real objects belonging to the companion real space. This limitation diminishes the sense of telepresence since the person perceives the technological medium and its constraints. Given this scenario, the question arises: "How can we empower avatars to physically move objects, such as 'moving the teapot', rather than relying on the friend to perform this task?"

In our previous work [5], we laid the definition of 'Physical Metaverse'and the groundwork for this ambitious objective by proposing a preliminary version of the Avatarm, i.e., an avatar able to manipulate objects in the real environment of a remote user using a robotic arm which is hidden from view and replaced in the video stream with the hand of the avatar. In addition to coining this new concept, in [5] we presented a method for diminishing the visibility of the robotic arm within the shared environment. This was achieved by using the computer-aided design (CAD) model and the kinematics of the robot to determine the region of interest within the frame. A preliminary implementation demonstrated the potentialities of the Avatarm.

With this work, we aim to significantly advance the infrastructure for the Physical Metaverse by integrating several technological components that will provide an

[1] https://youtu.be/Uvufun6xer8

immersive and engaging user experience, including haptic perception, robot awareness, and overall system performance. A simplified representative scenario of our vision of the Physical Metaverse is depicted in Fig. 1.

In particular, regarding the control algorithm, in our previous work the robot was teleoperated exploiting an open-loop control, resulting in inaccurate tracking of the digital object. In this study, we implemented a closed-loop control algorithm to ensure precise alignment between real objects and their digital twins. The virtual hand used in [5] has been replaced with a newly designed one capable of realistic interactions with virtual objects, combined with a digital avatar that mimics the user's upper body movements. Another limitation of our previous implementation was the lack of feedback on the object held by the robot gripper. To solve this, we integrated a cutting-edge wearable haptic interface featuring hand closure tracking and force feedback capabilities, and we developed a strategy to simultaneously manipulate virtual objects, teleoperate the robotic gripper, and map the force exerted by the gripper on the object to the user's fingertips. Additionally, the invisibility of the robot in the augmented environment, which is intrinsic to the Avatarm concept, posed a potential risk of demanding target positions that are inaccessible to the robot. Therefore, we integrated vibrotactile feedback provided at the wrist to re-enable the user to be aware of the robot's workspace, notifying them whenever the robot approaches singularities.

In summary, the contributions to the framework for the Physical Metaverse presented in this work include:

- integration of grasping force feedback;
- integration of situational awareness feedback;
- implementation of a closed-loop robot control algorithm;
- design of a novel virtual hand for more realistic manipulation of virtual objects;
- integration of a digital avatar in the augmented environment;
- a comprehensive experimental campaign aimed at validating the entire system, assessing each component's performance, ensuring the effectiveness and reliability of the proposed system, and evaluating the user experience.

The implementation of such a complete and complex system advances the state of the art by enabling the exploration of a new research field where robotics becomes functional in developing new platforms for realistic social interactions at a distance.

The rest of the paper is structured as follows. Section II provides the reader with a comprehensive literature review, while in Section III we report a detailed description of each building block of the Avatarm. In Section IV, we detail the implementation and validation of the Avatarm. This includes experiments specifically designed to assess both system performance and user experience. A discussion of the results and limitations is provided in Section V. Possible subjects of future research and conclusions are outlined in Section VI and Section VII, respectively.

II. RELATED WORKS

Virtual and augmented realities are currently primarily employed to visually modify or enrich the appearance of the physical environment. However, the full potential of digital realities to tangibly impact and reshape the real world has only been partially explored [6], [7]. In this direction, researchers investigated methods to achieve physical-virtual interactivity in XR within the scope of gaming experiences [8]. Despite the promising results, these gaming setups did not allow for object manipulation and were quite limited in their applications. Indeed, the proposed solutions have not fully integrated both virtual and physical worlds, failing to harness the two environments to their full potential [9]. To address this limitation, we propose the use of a robotic arm to replicate the interactions occurring in the virtual environment within the physical world.

This approach introduces new challenges, particularly from the users' perspective. For example, operators often face occlusion issues when manipulating robots, either due to the layout of the environment or the robot's body obstructing the user's view [10]. Diminished Reality (DR) techniques [11] have been investigated in human-robot interaction scenarios to address reduced visibility. These studies typically focus on enabling users to visualize occluded areas behind the robot [12], without exploring the potential of utilizing DR in the metaverse.

Regarding human-robot interaction in social contexts, existing techniques have focused on using robotic avatars as immersive tools for telemanipulation in remote environments. Schwartz et al. [13] implemented the NimbRo avatar system, an anthropomorphic avatar robot featuring two robotic arms in a humanoid configuration, and two dexterous robotic hands with different capabilities. The human operator controls the avatar by means of two robotic arms fixed to his arms which transmit the joint positions to the avatar, and receives feedback through an operator station that offers both visual and auditory immersion. Although research in this area is promising and could lead to significant advancements in the near future, the current complexity and obtrusiveness of these robotic avatars hinder their easy implementation and use for physical-virtual interactivity in XR.

To overcome these limitations, the technique proposed in this work is based on a single robotic arm that is concealed and substituted with a digital avatar in the virtual environment. A similar approach was undertaken in [14], where the appearance of a telepresence robot was augmented by overlaying a remote user on its structure, enhancing the perceived sense of immersion and psychological presence. The user drove the robot using the thumbstick of the handheld motion controller, while the avatar's arm gestures were controlled through motion tracking on the handheld controllers. This kind of setup provides the remote user with an identifiable self-embodiment and allows the local user to see the remote user's head direction and arm gestures. However, compared to our solution for the Physical Metaverse, the remote user does not have

manipulation capabilities and can only observe the remote environment.

The experienced psychological sense of presence [15], defined a s t he p erceptual i llusion o f n on-mediation i n the communicative environment [16], is particularly relevant because it can be considered an index of success of the metaverse. Fostering presence has been of particular interest to researchers as it influences t he i ntensity o f emotions felt in the virtual environment [17], [18], [19] and leads users to act as though the technological medium is non-existent [20]. A step forward is conveying two additional types of presence, i.e. the sense of co-presence and the sense of social presence. According to Youngblut [21], co-presence is defined a s ‘ ‘the s ubjective e xperience o f b eing together with others in a computer-generated environment, even when participants are physically situated in different sites’’. Social presence goes beyond co-presence by addressing also the social psychological idea of personal interaction. In line with Biocca [22], Youngblut proposed the following definition: ‘ ‘social p resence o ccurs w hen u sers f eel t hat a form, behaviour, or sensory experience indicates the presence of another individual. The amount of social presence is the degree to which a user feels access to the intelligence, intentions, and sensory impressions of another’’. Co-presence and social presence are fostered by the transmission of in-person interaction components when undertaking intentional, collaborative, or cooperative actions [23] including body language and physical contact.

In addition, research has demonstrated that collaborative interaction with shared objects enhances the overall quality and effectiveness of communication in the digital domain [24], [25], [26], [27]. In this context, haptic technology has been proven to be a powerful tool for enriching manipulation experiences in virtual reality and teleoperation settings [28], [29].While kinesthetic devices offer greater realism and accuracy in terms of force, cutaneous feedback is preferred in scenarios requiring wearability and portability. However, in the current state-of-the-art of wearable devices, the simultaneous integration of both hand tracking and force feedback has rarely been implemented. In this regard, Pierce et. al [30] designed a haptic device to control the parallel-jaw gripper of a remote robot, providing kinesthetic and cutaneous feedback to convey measurements of the grasping action. This device primarily facilitates easy tracking of pinch grasps, but does not allow for free manipulation of virtual objects, making it less suitable for virtual reality settings.

More in line with usage in XR is the TouchDIVER (WEART srl, IT) [31], a wearable device developed to simultaneously provide three types of feedback (pressure, vibration, and temperature) at the fingertips, along with integrated finger tracking and seamless hand tracking integration. While the TouchDIVER has been used for tactile feedback in both simulation [32] and real-world scenarios [33], as well as for hand tracking in virtual reality [34] and teleoperation of robots in assembly tasks [35], it has not yet been fully integrated into both teleoperation and virtual reality domains simultaneously with an HMD. In our work, we integrate all these features to enable users to interact with virtual objects in the digital domain while at the same time teleoperating a robotic gripper that transmits back the sensed force to the user’s fingertips. In addition, in [32] and [35] the end-effector of the robot is controlled according to the motion of the hand. In contrast, in our approach the robot is controlled to achieve the overlapping between the real object grasped by the robot and the digital twin manipulated by the user.

In addition to providing force feedback when interacting with objects, haptics have also been explored for alerting users in critical situations during a mutual interaction with a robot [36]. Equipping users with wearable haptic interfaces has been identified as an essential feature for improving performance in robot collaboration scenarios, particularly for users who may not be conscious of the robot’s actual pose because they are focused on completing a task effectively [37]. In our work, where the robot is invisible from the user’s perspective, including a haptic interface that alerts the remote user when the robot approaches configurations with reduced dexterity is essential for smooth teleoperation.

III. THE AVATARM

This section provides a comprehensive description of the Avatarm, starting with an application scenario designed to familiarize the reader with our vision, followed by a detailed explanation of the technical methodology for its implementation. Tables of all acronyms and symbols used here and in the subsequent sections are in Tab. 5 and Tab. 6 in Appendix.

A. APPLICATION SCENARIO

To facilitate the understanding of how we transform the avatar into the Avatarm, from now on we will take advantage of the scenario represented in Fig. 2 for presenting, explaining, and characterizing a possible scenario taking place in the Physical Metaverse.

Let us consider two people, Alice and Brad, having a conversation in a metaverse. The shared environment is an augmented version of Alice’s physical surroundings. They are sitting at the kitchen table having tea together at a distance. Alice’s kitchen is equipped with a camera and a robotic arm, which is remotely controlled by Brad and is able to manipulate objects on the table. This makes the robotic arm the ‘Brad-arm’, allowing Brad to experience Alice’s environment through an Avatarm. The resulting shared environment is therefore Alice’s kitchen augmented with the avatar of Alice (since she does not need to interact with distant objects) and the Avatarm of Brad, with the robotic arm hidden from view. Both Alice and Brad see the shared environment rendered in their HMDs, but from different points of view. Alice uses the camera mounted on her HMD, while Brad uses an extra camera placed in front of the table in Alice’s kitchen, where Brad would be sitting.

Now, let us suppose Brad wants to pour Alice a cup of hot water or pass her a biscuit. When Brad grasps an object

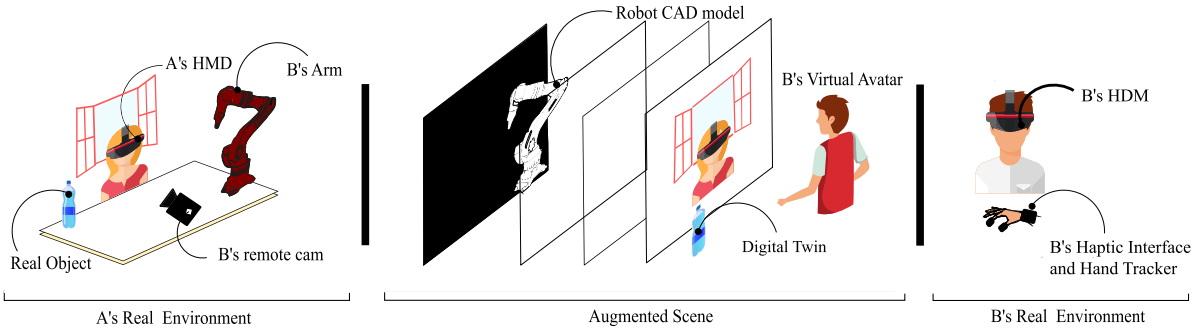

FIGURE 2. Illustration of a possible scenario in the Physical Metaverse. Two users, Alice (A) and Brad (B), are physically remote and participate in a metaverse session set in A's physical environment. Brad is the Avatarm of the scene. In the real world (left and right panel), a robotic manipulator and a manipulable object are placed on A's desk, a camera is placed where B would be sitting, and B's hand movements are acquired using a hand tracking system. In the virtual world (middle panel), the manipulator is removed from the scene and replaced with B's virtual avatar, and the digital twin of the manipulable object is added to the environment.

on Alice's table, the interaction between Brad's virtual hand and the physical object is managed through a digital twin of the object. Both Alice and Brad see Brad's virtual hand grasping and manipulating the digital twin of the object, while the real object is simultaneously manipulated by the robotic arm. To let the system work, the hand of Brad controls the virtual hand. The robotic arm feeds back its pose to enable the real-time overlay of its structure, as well as force data sensed at the end-effector when the object is grasped. Thanks to a wearable haptic interface, Brad perceives the real object during manipulation and receives haptic feedback when the robot approaches kinematic singularities. The latter feature is necessary to give Brad awareness of the space the robot can reach as he cannot see it.

B. METHODOLOGY

The aforementioned scenario is realized in this work through the integration of several technical modules, which are detailed in what follows.

1) ROBOT CONCEALMENT

The proposed framework is built upon the idea of performing the real-time concealing of the robotic arm to foster the embodiment and the realism of the interaction. This is done by developing a virtual environment incorporating a digital twin of the robot. The latter is implemented using the CAD model of the robot, ensuring that its applicability can be generalized to any manipulator, and robustifying our approach with respect to issues that usually affect DR based on image recognition [38]. For the sake of clarity, we will briefly report the key steps of the concealment method proposed in [5] and visually summarized in Fig. 3.

The first step consists in creating a digital twin of the robotic arm by exploiting the CAD models of its links and joints. This way, the digital twin can reproduce the kinematics of the robot in the virtual world using the real values of joints, measured during the task execution.

The digital twin is white-colored and placed in front of a black panel to facilitate the generation of the binary mask that identifies the portion of the scene covered by the robotic arm. A virtual camera captures the scene from a viewpoint mirroring the real camera perspective on the actual robot. Camera resectioning is done through an automatic procedure that utilizes multiple images of a calibration pattern (a chessboard) to estimate the intrinsic and extrinsic parameters of the physical camera, along with any potential lens distortions. The extrinsic parameters are estimated with respect to the base of the real robot and set as the relative position and orientation of the virtual camera with respect to the robot digital twin to ensure the alignment of the two points of view. The image frames of the videos acquired by the digital camera are expressed as time-varying 3D matrix $F(t)$ and are encoded using RGBα values. Pixels $\wp$ are distributed according to the coordinate system (x, y) within the size of the image (X, Y). In order to recognize the digital manipulator from the virtual camera, an image segmentation algorithm [39] is applied. The latter determines whether a pixel in $F(t)$, denoted as $\wp_F(x, y, t)$, contains a portion of the robot based on whether its RGBα values fall all within specific thresholds. The minimum and maximum values of the color and matte channels are chosen in accordance with the white color of the digital twin and the black color of the background panel. Subsequently, the binarization algorithm creates a mask $M(t)$ whose pixels $\wp_M(x, y, t) = 1$ if pixels $\wp_F(x, y, t)$ contain a portion of the robot image, and $\wp_M(x, y, t) = 0$, otherwise. The binary mask $M(t)$ is then combined with the color components obtained from a previous photograph of the background, taken in the absence of the robot. This combination produces a chromatic mask denoted as $H(t)$.

Image frames $I(t)$ captured by the real camera are transmitted into the virtual environment and rendered as dynamic textures on a virtual plane panel. The superimposition of the chromatic mask $H(t)$ onto the images $I(t)$ produces the resulting image $\tilde{I}(t)$, where each pixel $\wp_{\tilde{I}}$ is determined by:

$$\wp_{\tilde{I}}(x, y, t) = \begin{cases} \wp_I(x, y, t) & \forall x, y : \wp_M(x, y, t) = 0; \\ \wp_H(x, y) & \forall x, y : \wp_M(x, y, t) = 1. \end{cases}$$

At this stage, since the pixels within the mask are replaced with pixels of the background image, the robotic arm does not

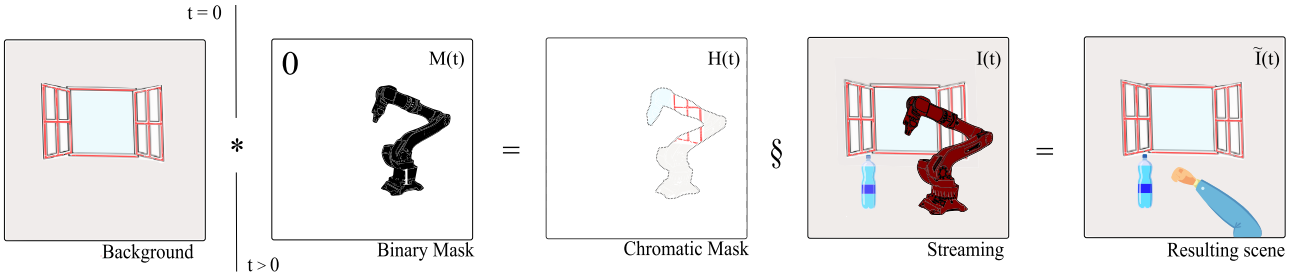

FIGURE 3. The robot concealment is performed by creating and applying a dynamic chromatic mask on the video streaming of the environment. Firstly (at $t = 0$), a photo of the background without the robot is acquired. At each time instant $t > 0$, the binary mask $M(t)$ is computed using the CAD model of the robot, and then combined with the background image to obtain the chromatic mask $H(t)$. The superimposition operation (denoted with the § operator) of the chromatic mask on camera frames $I(t)$ returns a scene diminished of the robot $\tilde{I}(t)$.

appear anymore in $\tilde{I}(t)$. The resulting image diminished of the robot is overlaid with the virtual twin of the manipulated real object and the digital avatar of the Avatarm, and eventually streamed into the HMD.

To ensure the digital twin of the object is manipulable, it should have both the same aspect and the same size of the real object, as well as maintain the same relative position with respect to the camera to prevent any potential misalignments. The main challenge in achieving this arises from the fact that our scene is projected onto a two-dimensional panel, while the object motion is performed within a three-dimensional virtual space. This discrepancy may cause the virtual object to disappear behind the panel when moved away from the point of view. To address this problem, we adjust the relative position of the digital object so that its dimensions match those of the real object only when the virtual object is positioned in front of the virtual layer, far enough to avoid it crossing the panel during the interaction. Then, the position of the virtual twin center and size are evaluated following the perspective transformation theory to match the pixels of $I(t)$ that contain portions of the real object, as mentioned in [40].

2) ROBOT MOTION CONTROL

There are mainly two approaches to address the problem of controlling the robot for the purposes of the Avatarm. One is to implement a control law that requires the robot end-effector to follow the hand of the user at each time instant, regardless of the particular action. Although this solution appears to be the simplest, it is actually not efficient as it implies the robot to follow the user even when it is not necessary (e.g., the person is gesturing and is not going to fetch an object). Moreover, it does not provide a general telemanipulation framework accounting for asymmetric teleoperation systems (e.g., the configuration of the end-effector is different from the human hand), thus it does not guarantee that the real object undergoes the same effects applied to the virtual object. The second solution is to combine the robot control with machine learning techniques that give the robot the ability to anticipate the human action by understanding which is the target object. This way, the robot control can be optimized to ensure that the end-effector approaches the object only when it is probable that the user is about to grab it. Once the object is grasped, the control law requires the robotic arm to reproduce on the real object the movement applied to the virtual object.

Between these two solutions, we consider the second one as the most promising long-term strategy. However, we opted not to implement the anticipation of human actions, as this feature is currently deemed irrelevant in this stage of the work. Indeed, this is not a novel strategy in robotics, as several works already demonstrated its feasibility [41], [42], [43], [44], [45], [46]. On the contrary, implementing the control law to map the motion applied on the digital twin of the object to the physical object is crucial to showcase the capabilities of the Avatarm.

In our previous work [5], the velocities at the center of mass of the digital object were directly mapped to the velocities of the real object in an open-loop manner, without any feedback action based on the real pose of the real object, leading to an unbounded escalation of the error. In this work, we implemented a closed-loop strategy to compute the joint velocities of the robot taking into consideration the pose of the real object. A diagram representing the flow of information for the robot motion control is in Fig. 4.

The reference for the robot control is the pose of the center-of-mass of the digital twin of the object o in the virtual reference frame $\mathbb{V}$, ${}^{v}\varrho_o(t) = \left[{}^{v}p_o(t)^T \; {}^{v}\theta_o(t)^T\right]^T$, where the two vectors ${}^{v}p_o \in \mathbb{R}^3$ and ${}^{v}\theta_o \in \mathbb{R}^3$ represent position and orientation, respectively. At each time instant t, ${}^{v}\varrho_o(t)$ is acquired and filtered with a moving average filter (MAF) to compensate all possible jitter variations due to the interaction between the virtual object and the virtual hand, i.e.:

$$ {}^{v}\hat{\varrho}_o(t) = \begin{cases} \dfrac{1}{\delta t}\displaystyle\int_{t-\delta t}^{t} {}^{v}\varrho_o(\tau)\delta\tau & \text{if } t \geq \delta t \\ {}^{v}\varrho_o(t) & \text{otherwise} \end{cases} $$

where δt is the MAF time frame.

Let ${}^{w}T_o \in \mathbb{R}^{4\times4}$ and ${}^{v}T_o \in \mathbb{R}^{4\times4}$ be homogeneous transformations expressing the object pose in the world reference frame $\mathbb{W}$ and in $\mathbb{V}$, respectively. Then, the filtered object pose can be expressed in $\mathbb{W}$ according to the following equation:

$$ {}^{w}T_o(t) = {}^{w}\tilde{T}_v \, {}^{v}T_o(t) \tag{1} $$

where

$$^{w}\tilde{T}_{v} = {}^{w}T_{\hat{v}}\ {}^{\hat{v}}A_{v}$$

transforms a point expressed in $\mathbb{V}$ into a point expressed in $\mathbb{W}$ through the matrix ${}^{\hat{v}}A_{v}$ and the homogeneous transformation matrix ${}^{w}T_{\hat{v}} \in \mathbb{R}^{4\times4}$. The matrix ${}^{\hat{v}}A_{v}$ is needed to account for other mapping actions in an auxiliary reference frame $\hat{\mathbb{V}}$ (e.g., projection of a left-handed reference frame into a right-handed one). Analogously to Eq. (1), the desired pose of the end-effector at time t expressed as the transformation matrix ${}^{w}T_{e}(t) \in \mathbb{R}^{4\times4}$ is:

$$^{w}T_{e}(t) = {}^{w}T_{o}(t)\,{}^{o}T_{e}$$

where ${}^{o}T_{e}$ is the homogeneous transformation mapping the end-effector frame to the object frame. The latter is computed as:

$$^{o}T_{e} = {}^{w}\bar{T}_{o}^{-1}(t_{\Gamma})\,{}^{w}\bar{T}_{e}(t_{\Gamma})$$

being ${}^{w}\bar{T}_{o}$ and ${}^{w}\bar{T}_{e}$ the pose in $\mathbb{W}$ of the end-effector and of the object measured at the time instant t_{Γ}, i.e. , when the object is grasped. This formulation holds under the hypothesis that the grip is firm and there is no sliding between the object and the gripper, meaning that the relative position and orientation between the object and the end-effector do not change after t_{Γ}.

To compute the reference signal for the end-effector velocity, firstly the error between actual and desired end-effector pose expressed in form of homogeneous transformation matrix ${}^{w}\Delta T_{e}(t)$ is calculated as:

$$^{w}\Delta T_{e}(t) = \begin{bmatrix} {}^{w}\Delta R_{e}(t) & {}^{w}\Delta p_{e}(t) \\ 0_{1\times3} & 1 \end{bmatrix}$$

being

$$^{w}\Delta R_{e}(t) = {}^{w}R_{e}(t)\,{}^{w}\bar{R}_{e}^{-1}(t)$$
$$^{w}\Delta p_{e}(t) = {}^{w}p_{e}(t) - {}^{w}\bar{p}_{e}(t)$$

the orientation and position errors. With ${}^{w}R_{e}$ and ${}^{w}\bar{R}_{e}$, and ${}^{w}p_{e}$ and ${}^{w}\bar{p}_{e}$ we denote desired and real end-effector orientations, and desired and real end-effector positions, respectively. Then, the desired end-effector velocity is obtained as:

$$^{w}\nu_{e}(t) = \begin{bmatrix} {}^{w}\dot{p}_{e}(t) \\ {}^{w}\omega_{e}(t) \end{bmatrix} = \begin{bmatrix} \Delta t^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Delta t^{-1} \end{bmatrix} \begin{bmatrix} {}^{w}\Delta p_{e}(t) \\ {}^{w}\Delta\theta_{e}(t) \end{bmatrix}$$

where Δt^{-1} corresponds to the update frequency of the virtual scene (the update rate of ${}^{v}\varrho_{o}$), and ${}^{w}\Delta\theta_{e}(t) \in \mathbb{R}^{3}$ is the end-effector orientation error in Roll-Pitch-Yaw angles form. After the computation, the desired linear velocity of the end-effector is saturated with a threshold value of $0.25\ \mathrm{ms}^{-1}$ to comply with the ISO-10218 standard for human-robot interaction [47], obtaining the ratio:

$$r_{\dot{p}}(t) = \frac{\|{}^{w}\dot{p}_{e}(t)\|}{0.25\ \mathrm{ms}^{-1}}.$$

The resulting reference signal for the end-effector velocity is:

$$^{w}\tilde{\nu}_{e}(t) = \Psi^{-1}(t)\,{}^{w}\nu_{e}(t)$$

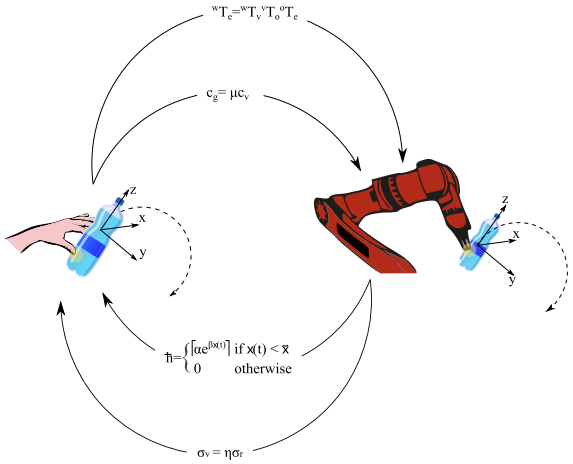

FIGURE 4. Information exchange between the user controlling the Avatarm and the robot. The pose of the center-of-mass of the object digital twin is used as reference for the pose of the end-effector. The closure state of the virtual fingers is mapped into the closure state of the gripper, while the internal forces measured by the sensors placed on the jaws are fed back to the real hand, together with the manipulability measure of the robot.

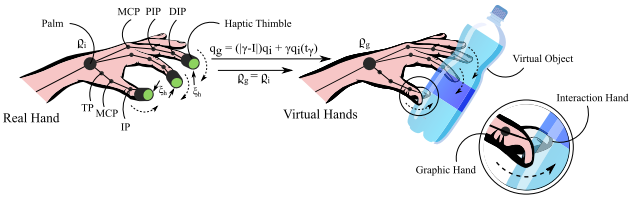

FIGURE 5. Information exchange between the user controlling the Avatarm and the virtual hands. When no contact with digital objects occurs, the palm pose of the user's hand is consistently mirrored by both the (hidden) interaction hand and the graphical hand. Whenever a contact occurs, the graphical hand remains in the position where the contact occurred, while the interaction hand continues to track the user's movements in order to compute feedback forces according to the mapping strategy.

with

$$\Psi(t) = \begin{cases} I_{6\times6} & \text{if } r_{\dot{p}} \leq 1 \\ diag([r_{\dot{p}}, r_{\dot{p}}, r_{\dot{p}}, I_{3\times3}]) & \text{otherwise.} \end{cases}$$

Finally, the reference joint velocities $\dot{q}(t)$ are computed exploiting the pseudo-inverse[2] of the Jacobian matrix $J(q(t))$ of the robot:

$$\dot{q}(t) = J^{\dagger}(q(t))\,{}^{w}\tilde{\nu}_{e}(t).$$

3) AVATARM HAND

In our framework, we replace the robot manipulator with a virtual avatar that mimics the user's commands. To ensure a reliable physical interaction in the digital environment, we customized the avatar hands. Specifically, the avatar's hand should be capable of realistically interacting with virtual objects.

The complete human hand model has about 30 degrees of freedom (DoFs) [48]. In this work, we used a simplified

[2]We use the operator $(\cdot)^{\dagger}$ to denote the pseudoinversion operation

26 DoFs kinematic model to reconstruct a digital hand in virtual space. We modelled the fingers as 4-DoF kinematic chains to computationally reduce the rendering costs while preserving the biomechanics of the human hand [49]. In particular, each finger has 2 DoFs at the metacarpophalangeal (MCP) joint for the abduction/adduction and flexion/extension of the metacarpal bone, 1 DoF at the proximal-interphalangeal joint, and 1 DoF at the distal-interphalangeal joint. As regards the thumb, we modelled it with 1 DoF at the trapeziometacarpal joint, 2 DoFs at the MCP joint, and 1 DoF at the interphalangeal joint. The resulting virtual hand model has 26 DoFs, i.e. 20 DoFs describing the fingers and 6 DoFs for the palm position and orientation. Palm pose and joint angles of the real hand are acquired at each time instant by a tracking system, and assigned to palm pose $\varrho_i(t) \in \mathbb{R}^6$ and joint angles $q_i(t) \in \mathbb{R}^{20}$ of the virtual hand, as visually depicted in Fig. 5.

The virtual hand is designed as a composition of transparent capsules, and the contact between the j-th finger and the object is detected considering the volume of their intersection. More in detail, for each finger j we compute the volume of the intersection between each phalanx and the object. We denote the sum of these values as V_j. If $V_j = \emptyset$, then the finger j is not in contact with the object; otherwise, the finger is touching the object. To prevent the user from seeing the virtual hand penetrating the objects when these are grasped, we superimpose a second virtual hand on the previous one. For the sake of clarity, from now on we will refer to the first virtual hand as *interaction hand*, while the superimposed virtual hand will be indicated as *graphic hand*. This solution enables the user to command a greater closure of the gripper compared to the size of the object, thereby ensuring that the gripper applies the necessary force to grasp the object, all while avoiding inconsistencies in the virtual environment.

Palm pose ϱ_g and joint angles q_g of the graphic hand coincide with ϱ_i and q_i of the interaction hand until the contact with the virtual object is detected. After the time instant of the first contact t_γ, the joint angles update of the graphic hand is stopped, hence:

$$\begin{cases} \varrho_g(t) = \varrho_i(t) \\ q_g(t) = (|\gamma(t) - I_{20\times 20}|)q_i(t) + \gamma(t)q_i(t_\gamma) \end{cases}$$

where $\gamma(t) \in \mathbb{R}^{20\times 20}$ is the diagonal matrix whose submatrices $\gamma_j(t) \in \mathbb{R}^{4\times 4}$ depend on the contact state of the j-th finger of the interaction hand, i.e.

$$\gamma(t) = \begin{bmatrix} \ddots & & \\ & \gamma_i(t) & \\ & & \ddots \end{bmatrix} \text{ with } \gamma_i(t) = \begin{cases} I_{4\times 4} \text{ if } V_j \neq \emptyset \\ 0_{4\times 4} \text{ if } V_j = \emptyset. \end{cases}$$

a: GRIPPER CLOSURE CONTROL

While the virtual object pose is used as reference for the end-effector positioning, the gripper closure state $c_g(t) \in \mathbb{R}$ is controlled using the information about apposition/opposition of the virtual fingers of the Avatarm. In particular, we evaluate the closure state of the virtual interaction hand $c_v(t) \in \mathbb{R}$ as distance in ℓ_2-norm between the position of the thumb fingertip, $p_{i,t}(t) \in \mathbb{R}^3$, and the average position of the other fingertips, denoted as $p_{i,\phi}(t)$ with $\phi \in [1, \ldots, n]$ as we consider only the n tracked fingers. Thus,

$$c_v(t) = \|p_{i,t}(t) - \frac{1}{n}\sum_{\phi=1}^{n} p_{i,\phi}(t)\|_2.$$

A scaling factor μ is applied to c_h to account for differences in maximum reachable distance between the fingers and the gripper. Moreover, to avoid damages at the end-effector, the force applied by the gripper is controlled to not exceed the maximum safe force applicable Λ_r. Hence, the gripper closure state is given by:

$$c_g(t) = \begin{cases} \mu c_v(t) & \text{if } \lambda_r < \Lambda_r \\ c_g(t - \Delta t) - \delta c_g & \text{if } \lambda_r \geq \Lambda_r \end{cases}$$

where $\lambda_r = \begin{bmatrix} f_{r,1} & \tau_{r,1} & \ldots & f_{r,m} & \tau_{r,m} \end{bmatrix}^T$ is the vector of the generalized forces applied at the m contact points. This formulation implies that the jaws start to open with a fixed step δc_g if one or more components of the contact forces exceed the safety limits, reaching step-by-step a more safe closure until the sensed forces are less than Λ_r.

b: FORCE FEEDBACK

The squeezing forces applied on the object by the robot are fed back to the user who controls the Avatarm through a wearable haptic interface.

The force feedback is computed starting from the generalized force λ_r applied by the jaws of the end-effector on the object, which is measured by the sensors mounted on the gripper. According to the grasping theory [50], the wrench applied to the gravity center of the object $w_r = \begin{bmatrix} f_r & \tau_r \end{bmatrix}^T \in \mathbb{R}^6$ is computed as

$$w_r(t) = G_r(t)\lambda_r(t)$$

where $G_r(t) \in \mathbb{R}^{6\times 6m}$ is the grasp matrix defined as follows:

$$G_r(t) = \begin{bmatrix} P_{r,1}\overline{R}_{r,1} & \ldots & P_{r,m}\overline{R}_{r,m} \end{bmatrix}.$$

The matrix $P_{r,j}$ defines how variations in the position and orientation of the j-th contact point, expressed in the reference frame of the contact point, affect the object center of gravity:

$$P_{r,j} = \begin{bmatrix} I_{3\times 3} & 0_{3\times 3} \\ S(p_{r,j}(t) - \hat{p}_r(t)) & I_{3\times 3} \end{bmatrix}.$$

Here, $S(p_{r,j}(t) - \hat{p}_r(t))$ represents the cross product matrix of the distance between the position of the j-th contact point, denoted as $p_{r,j}$, and the position of the gravity center, denoted as $\hat{p}_r(t)$. $\overline{R}_{r,j}$ is a block diagonal matrix describing the orientation of the j-th contact frame with respect to the inertial frame $R_{r,j}$, that is:

$$\overline{R}_{r,j} = \begin{bmatrix} R_{r,j} & 0_{3\times 3} \\ 0_{3\times 3} & R_{r,j} \end{bmatrix}.$$

The squeezing force $\xi_r \in \mathbb{R}^{3m}$ on the real object is defined as the vector consisting of forces $\xi_{r,j}$ applied at the j-th contact point that do not contribute to w_r:

$$\xi_r(t) = K_r \underbrace{\left(I_{6m\times 6m} - G_r^{\dagger} G_r\right)}_{\mathcal{N}(G_r)} \lambda_r(t)$$

where $\mathcal{N}(\cdot)$ indicates the null space of a generic matrix, while $K_r \in \mathbb{R}^{3m\times 6m}$ is a selection matrix used to extract only the force components.

The squeezing forces are mapped from the gripper to the user fingertips in accordance with the backward mapping procedure described in [51]. Firstly, the magnitude of the squeezing forces at the gravity center of the real object $\sigma_r(t) \in \mathbb{R}$ is computed as the sum of the ℓ_2-norms of the internal forces for each contact point j:

$$\sigma_r(t) = \sum_{j=1}^{m} \|\xi_{r,j}\|_2.$$

$\sigma_r(t)$ is mapped into the squeezing force at the gravity center of the virtual object $\sigma_v(t) \in \mathbb{R}$ following the equation:

$$\sigma_v = \eta \sigma_r$$

where

$$\eta = \frac{\Xi_h}{\Lambda_r}$$

is a scale factor equal to the ratio between the maximum force achievable by the actuators embedded in the haptic interface $\Xi_h \in \mathbb{R}$, and the maximum force Λ_r defined in Section III-B2. Then, by defining with λ_v the virtual forces applied at the n_c in-contact fingers of the interaction hand, magnitude and direction of the internal forces $\xi_h \in \mathbb{R}^{3n_c}$ to render through the haptic interface are given by:

$$\xi_h(t) = \begin{cases} \dfrac{K_v \mathcal{N}(G_v)\lambda_v(t)}{\|K_v \mathcal{N}(G_v)\lambda_v(t)\|} \dfrac{1}{(n_c)} \sigma_v(t) & \text{if } n_c \neq 0 \\ \emptyset & \text{if } n_c = 0 \end{cases}$$

where $K_v \in \mathbb{R}^{3n_c\times 6n_c}$ selects the force components, as for the real object. The components of ξ_h are digitally codified to control the haptic interface.

4) SITUATION AWARENESS

Since the robot is hidden from view in the Physical Metaverse, the user who controls the Avatarm is not aware of the boundaries of the reachable workspace of the manipulator. This means that the user could try to move an object to a point that the robot cannot reach, with obvious negative consequences on the functioning of the system. To cope with this possible problem, we make the user aware of the distance to singular configurations by mean of a vibrotactile signal provided through a haptic interface. This solution guarantees the functioning of the Avatarm without compromising the realism achieved with the real-time concealment of the robotic arm.

To compute the distance of the manipulator from singular configurations, we exploit the manipulability measure $\chi(t)$ of the robot, i.e.:

$$\chi(t) = \sqrt{\det(J(q(t))J(q(t))^T)}.$$

The corresponding vibrotactile signal is evaluated as:

$$\hbar(t) = \begin{cases} \left\lceil ae^{-b\chi(t)} \right\rceil & \text{if } \chi(t) \leq \tilde{\chi} \\ 0 & \text{if } \chi(t) > \tilde{\chi} \end{cases}$$

where $\tilde{\chi}$ is the lower bound of manipulability for situation feedback and the coefficients a and b are determined depending on the haptic interface adopted. It is worth noting that this formulation generates a vibration with a dead zone around the center of the robot workspace, avoiding the sensory overload of the user during the interaction in the Physical Metaverse.

IV. EXPERIMENTAL EVALUATION

To assess the feasibility of the Avatarm, we implemented the framework described in Section III and tested its performance under different aspects, considering both objective and subjective measures. In particular, the first two experiments were conducted to evaluate the capabilities of the algorithms for robot motion control and robot concealment (Section IV-C), and force feedback (Section IV-D), respectively. The last experiment (Section IV-E) was aimed at investigating the experienced psychological sense of presence, co-presence, and social presence given by the Avatarm.

A. IMPLEMENTATION

The experimental setup consisted of a Sawyer manipulator (Rethink Robotics GmbH, DE), two HMDs Oculus Quest 2 (Meta Platforms, Inc., US), and a TouchDIVER (WEART srl, IT).

The users surroundings were recorded using two full HD 1080p USB webcams (C920HD Logitech, US) with focal length 3.67 mm. The shared environment and the instrumental software layers were implemented using Unity Graphic Engine (Unity Technologies, US) and ROS. The binary mask $M(t)$ was computed using acquisitions from an 8-bit RGBα virtual camera observing the CAD-based model of the Sawyer. The uniformity thresholds on the RGBα space were set in accordance with the chromatic choices of the robot model and the color depth of the virtual point of view. For the camera resectioning procedures, we used the Camera Calibrator Toolbox for Matlab (MathWorks Inc, US). The relevant parameters of the system are reported in Tab. 1.

The user controlling the Avatarm joins the shared environment as a personalized avatar implemented using the Meta Avatars SDK, and integrated with the virtual hand described in Section III-B3. The user movements, including head and wrists, were tracked directly with the HMD and its controllers. A pre-built inverse kinematics algorithm mapped the motion of the devices to the pose of the avatar. The real hand closure was tracked using the TouchDIVER SDK,

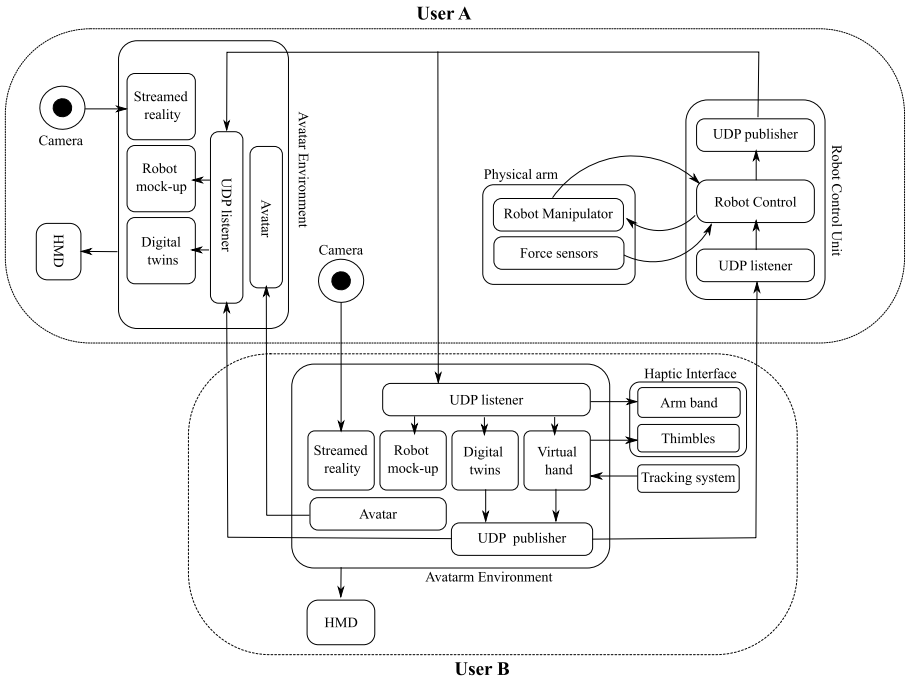

FIGURE 6. Block diagram representing the system interconnection for the Physical Metaverse.

which combines inertial data and vision-based acquisitions. The interaction with virtual objects was realized using the physical engine PhysX (NVIDIA Corporation, US). The Oculus controller mounted on the TouchDIVER was used to actuate the vibrotactile signal for providing user's situation awereness (Section III-B4).

The robot end-effector was a parallel-jaw gripper equipped with force sensors. Grasping forces were recorded using two ATI nano 17 sensors (ATI Industrial Automation, Inc., US), and fed back on the user's fingertips (thumb, index and middle fingers) through the thimbles of the TouchDIVER.

ROS Melodic on Ubuntu 18.04 was used for connecting locally all the system components, while the connection with the remote systems was realized using the User Datagram Protocol (UDP). The map of the system interconnection is visually depicted in Fig. 6.

B. PARTICIPANTS

Overall, 26 participants were enrolled in the study. The sample size was 10 in the first experiment (6 males, 4 females, age 36±12, all right-handed), 10 in the second experiment (6 males, 4 females, age 39 ± 20, all right-handed), and 20 in the third experiment (14 males, 6 females, age 36±12, all right-handed). Six participants of the second experiment participated also in the first experiment, eight participants of the third experiment also participated in the first experiment, while four participants took part in the entire experimental campaign.

Each participant gave their written informed consent to participate and was able to discontinue participation at any time during the experiments. The experimental evaluation protocols followed the declaration of Helsinki. Data were recorded in conformity with the European General Data Protection Regulation 2016/679, stored on local repositories with anonymized identities (e.g., User1, User2), and used only for the post processing evaluation procedure. No sensitive data were recorded.

TABLE 1. Parameters values used for the implementation of the Physical Metaverse.

Parameter	Description	Value
X	Length of images in pixels	960
Y	Height of images in pixels	720
δt	MAF time-window	1.00 s
m	Number of gripper jaws	2
$n+1$	Number of interactive fingers	3
μ	Hand to gripper scale factor	2.73
δc_g	Fixed step of gripper jaws	0.3 cm
$F_{r,j}$	Robot force limit	[20, 20, 20] N
$T_{r,j}$	Robot torque limit	[3, 3, 3] Nm
$\Xi_{h,f,j}$	Haptic interface force limit	[5, 0, 0] N
$\Xi_{h,\tau,j}$	Haptic interface torque limit	[0, 0, 0] Nm
$\tilde{h}$	Max digital value of haptic interface	1
$\tilde{\chi} = 0.1 \sup(\chi)$	Manipulability lower bound	0.1
$[a, b]$	Haptic signal parameters	$[1, -80]$

C. EXPERIMENT 1—ROBOT MOTION CONTROL AND ROBOT CONCEALMENT

This experiment aimed at evaluating both the motion accuracy of the robot, and the goodness of the algorithm for hiding the manipulator from view in the Physical Metaverse. While it was possible to conduct this experiment through

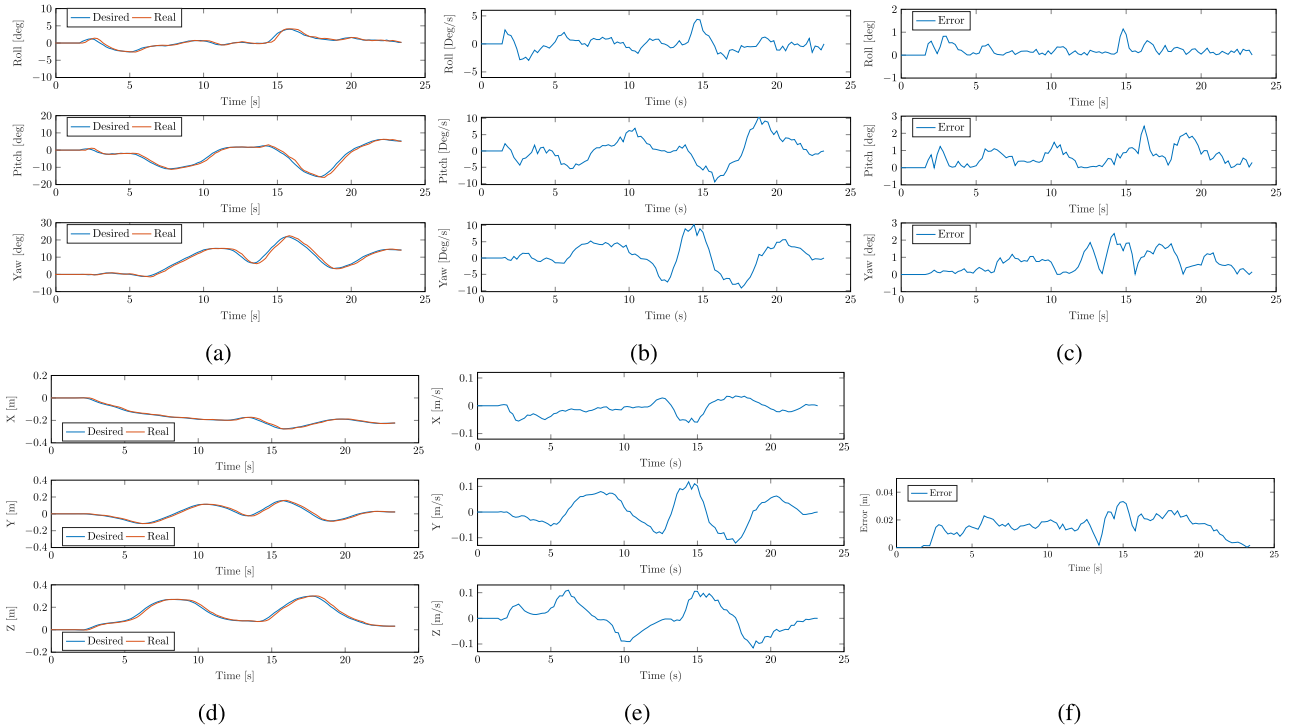

FIGURE 7. Experiment 1. Representative trial in which the user grasped the cube, lifted and moved it tracing two consecutive arcs in mid-air. In a), the orientation of the center-of-mass of the real cube and of the digital twin (expressed in RPY convention) are depicted in red and blue, respectively. In b), the angular velocity of the reference frame attached to the digital cube center-of-mass. In c), the RMSE between the desired and actual orientation. In d), the position of the center-of-mass of the real cube and of the digital twin are depicted in red and blue, respectively. In e), the linear velocity of the digital cube center-of-mass. In f), the Euclidean distance between the two centers-of-mass.

simulation, we opted to engage subjects in order to gather data from real use-cases.

Participants were comfortably seated on an office chair and were asked to wear the HMD and the TouchDIVER on their dominant hand. The proposed XR environment consisted of a remote environment with a desk and a rigid brown cube placed on it. As in the complete implementation of the Physical Metaverse, users could observe the scene diminished of the manipulator and augmented with their avatar and the digital twin of the cube, but no interlocutor was present in the scene, since it was not necessary for the purposes of the experiment. The task required each subject to pick and freely move the cube using the Avatarm. Each user moved and manipulated the object only once with no time constraints. The only indication given to the participants was not to throw the cube to avoid the safety lock of the robot. Overall, ten trajectories of different durations (average 25.3 s) were sampled at 5 Hz.

For each trial, we acquired both the pose of the center-of-mass of the cube digital twin, and the pose of the center-of-mass of the real cube. We considered these two measures as input and output signals for the purposes of the evaluation.

a: ROBOT MOTION CONTROL

Two different metrics were used for evaluating the robot motion control algorithm : *i)* the Root Mean Square Error (RMSE) between the desired (i.e., the input signal) and the actual (i.e., the output signal) object pose, and *ii)* the Normalized Root Mean Square Error (NRMSE) obtained normalizing the RMS error with respect to the maximum variation observed in each trial.

TABLE 2. Experiment 1. Average RMSE and NRMSE among the trials for position and orientation of the center-of-mass of the real cube with respect to the pose of the center-of-mass of the cube digital twin.

	RMSE	NRMSE
Position	1.27 ± 0.70 cm	$6.01\% \pm 2.33\%$
Roll	0.69 ± 0.77 deg	$3.88\% \pm 3.18\%$
Pitch	0.56 ± 0.69 deg	$2.42\% \pm 2.46\%$
Yaw	0.80 ± 0.35 deg	$1.88\% \pm 1.85\%$

The resulting average RMSE and NRMSE for position and orientation, along with the respective standard deviation, are reported in Tab. 2, while a representative trajectory is reported in Fig. 7 and Fig. 8. These results demonstrate the high accuracy of the robot motion control algorithm, as, among all trials, the average error in positioning the end-effector was less than 1.5 cm, with minimal average misalignment between the desired and real orientation (less than 1° on each axis). During the experimental evaluation, users manipulated and moved the virtual objects with a velocity having an average norm of 4.51 ± 2.69 cm/s, while the average rotational velocity of the virtual objects around

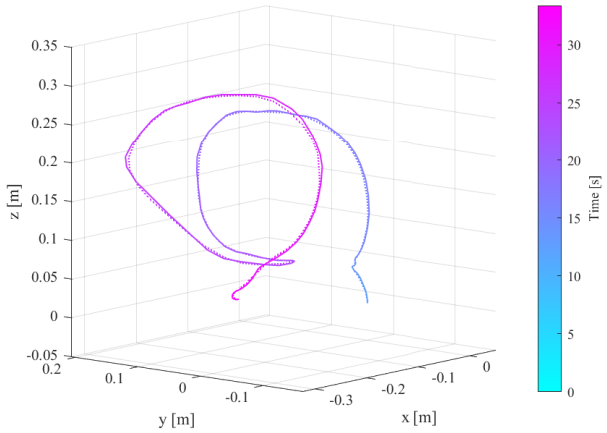

FIGURE 8. Experiment 1. Trajectories of the center-of-mass of the real cube and of its digital twin in a representative trial. The path of the digital twin is depicted with a dotted line, while the solid line depicts the path followed by the real cube.

x, y and z-axis were 4.06 ± 4.17 deg/s, 4.75 ± 4.97 deg/s, 3.13 ± 2.03 deg/s, respectively. These minor errors were primarily due to the latency of the overall system, since both Unity and ROS operate at a frequency of 5 Hz.

b: ROBOT CONCEALMENT

To assess the effectiveness of the algorithm without any biases due to covering part of the robot with the grabbed object or with the avatar, we used the input signal to play back the robot movement without the real cube in the gripper. For each trial we recorded two videos (with the same duration) at a frame rate of 30 frames per second, one of the real environment (i.e., before hiding the robot) and one of the environment diminished of the robot.

A Matlab-based software, developed for the purpose of the analysis, was used to compute the percentage of the robot concealed, in order to evaluate possible errors related to camera misalignment, system latency and CAD imperfections concerning the actual manipulator. In particular, for each frame the software counts the number of robot pixels for the two videos. Hence, the goodness of the concealing algorithm in a single trial is evaluated as:

$$\mathscr{M}_{\%} = \frac{1}{N}\sum_{k=1}^{N}\left(1 - \frac{\#\wp_{r,k}}{\#\tilde{\wp}_{r,k}}\right)\cdot 100$$

where N is the number of video frames of a single video, and $\#\wp_{r,k}$ and $\#\tilde{\wp}_{r,k}$ are the number of robot pixels for each frame k in the videos of the diminished and of the real environment, respectively. In the best case, the number of robot pixels in the video of the diminished environment is zero, thus the concealing success is 100%. Analysis of the recorded videos revealed that on average the algorithm was able to remove the 95.3% ± 3% of the manipulator. A comparison between two frames taken from a representative trial is in Fig. 9, while the video of a full trial is available online.[3]

FIGURE 9. Experiment 1. Comparison of frames after (left panel) and before (right panel) applying the robot concealment algorithm. The residual pixels that contain the robot after the concealment are red circled in the frame on the left.

FIGURE 10. Experiment 2. The user immersed in the augmented environment picks and places a soft yellow cube simulating a fragile object. The force data are acquired by the ATI sensors integrated on the robot gripper, and reproduced by the haptic interface on the fingers of the user.

D. EXPERIMENT 2—FORCE FEEDBACK

Similarly to the previous experiment., participants were asked to sit and wear the HMD and the TouchDIVER, and no interlocutor was present in the scene. The XR environment was the same proposed in the previous experiment. However, in this case the cube was yellow, soft, and with edges of 4 cm, and an additional cup was placed on the desk.

Participants had to use their Avatarm to pick the yellow cube and place it inside the cup (see Fig. 10). To simulate an interaction with a fragile object and create a challenging task, a grasping force threshold of 12 N was set, above which the object was considered broken, and thus the task failed. Whenever the force threshold was exceeded, the color of the cube digital twin shifted from yellow to red. Start and goal positions remained consistent across all experimental trials. Two distinct experimental conditions were implemented: with and without the haptic feedback. Before the start of the experiment, participants could familiarize themselves with the system by trying both the interaction modalities. Each participant repeated the task five times in a pseudo-random order for each experimental condition. Performance was evaluated using two metrics: the number of successes in accomplishing the task, and the impulse of the force. A video of the experiment is publicly available.[4]

Out of a total of 100 trials (50 with haptic feedback and 50 without), users failed to accomplish the task 52 times. Specifically, 19 of these failures occurred while perceiving haptic feedback, while the remaining 33 occurred without haptic feedback. For what concerns the failures with haptic feedback, the object was broken 14 times and dropped

[3]https://youtu.be/r6AyrwbPaF0

[4]https://youtu.be/vNpJuX9sOkU

5 times. Regarding the trials without feedback, the object was broken 28 times and dropped 5 times.

A paired-samples t-test was conducted to determine whether there was a statistically significant mean difference between the number of successfully accomplished tasks with and without the feedback. The assumption of normality was not violated, as assessed by Shapiro-Wilk's test ($p = 0.732$). Participants scored a higher percentage of successes when the haptic feedback was provided (62.0% ± 39.4%) as opposed to controlling the Avatarm with no haptic feedback (34.0% ± 25.0%). The test revealed a statistically significant increase of 28.0% of success rate (95% CI, 7.6% to 48.4%), $t(9) = 3.096$, $p = 0.013$.

A deeper analysis was carried out to determine the primary cause of failure, whether it was due to the dropping or breaking of the objects. Two further paired-samples t-tests were run to assess whether there was a statistically significant mean difference between the number of drops and breakages having or not the haptic feedback. For what concerns the breakages, Shapiro-Wilk's test assessed the assumption of normality ($p = 0.731$). A statistical significant reduction of 28% failures was observed when using the haptic feedback (28.0% ± 25.3%) with respect to trials performed with no feedback (56.0% ± 29.5%), 95% CI from 7.54% to 48.45%, $t(9) = 3.096$, $p = 0.013$. Conversely, the test on the drops revealed no statistical significance on the mean difference. This outcome was expected, as users prioritized a firm grasp, even at the risk of breaking the object, over the possibility of it falling.

The second metric (i.e., the impulse) was used for supporting the outcomes of the first metric. The impulse takes into account both force and time taken to accomplish the task, under the assumption that a lower impulse indicates a more successful execution of the experiment [52]. To this end, we analyzed only the successful trials, discarding the ones with broken or dropped objects. 31 trials were accomplished with force feedback and 17 without. No outliers were identified in the data, but the assumption of homogeneity of variances was violated (Levene's test for equality of variances $p < 0.001$). Thus, a Welch t-test was conducted to determine if there were differences in impulse between the two cases. Impulses were normally distributed, as assessed by Shapiro-Wilk's test ($p > 0.05$). The test revealed that the impulse was smaller with haptic feedback (88.06 N s±42.02 N s) than with no feedback (156.94 N s ± 138.79 N s), with a statistically significant difference of 68.88 N s (95% CI, 6.36 to 144.12), $t(16.58) = 1.935$, $p = 0.047$.

E. EXPERIMENT 3—USER EXPERIENCE

Finally, to assess user experience in a possible scenario of the Physical Metaverse, we conducted an experiment in the form of a collaborative and competitive game between teams of two participants. The experiment's purpose was to evaluate presence, co-presence, and social presence senses, comparing the experience of the Avatarm with respect to the one of the common avatar.

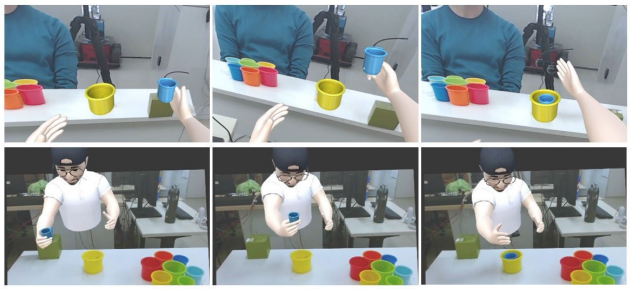

FIGURE 11. Experiment 3. Frames acquired during a representative trial from the point of view of the user controlling the Avatarm (top panels) and the point of view of the user controlling the avatar (bottom panels).

TABLE 3. Experiment 3. Teams' total scores. The maximum score achievable was 48.

Team	1	2	3	4	5	6	7	8	9	10
Score	18	39	38	27	42	38	34	30	20	18

The goal of the collaborative and competitive game was to cooperate with one partner to stack up to nine cups of decreasing diameter with the aim of outperforming other teams. In a series of six repetitions of the task, each pair of participants took turns in controlling the avatar and the Avatarm, switching roles after every three trials.

The shared environment for the experiment consisted of a room with a desk upon which nine cups (taken from the Yale-CMU-Berkeley Object and Model set [53]) were placed.

The scene was diminished of the manipulator, and augmented with the avatar of the user controlling the Avatarm and the digital twin of the smallest cup (see Fig. 11). The largest cup had a diameter of 8.5 cm and was pre-located on the desk to be used as a reference for positioning the other items. The participant controlling the avatar was asked to wear the HMD and seat in front of the desk. The participant controlling the Avatarm wore the HMD and the TouchDIVER, and was seated in a different room. Audio communication was provided by voice call.

In each trial, the Avatarm was asked to place the smallest (and the last) cup, which had a diameter of 4.8 cm, while the avatar managed the remaining eight cups in accordance with the partner instructions. The more cups were placed, the more challenging the task became for the Avatarm. The trial score was determined by adding the number of cups successfully stacked, ranging from 0 to 8. If the user controlling the Avatarm failed to place the final cup, the trial score was recorded as 0. The team total score was determined by summing all trial scores, promoting communication and strategic planning among the team. Scores are reported in Tab. 3. A video of the experiment is publicly available.[5]

As previously stated (see Section I), the experienced psychological sense of presence, co-presence, and social presence can be used to measure the success of the metaverse, thus they can be exploited as metrics for evaluating the soundness of our framework. Hence, after the first three trials

[5]https://youtu.be/sO0WCDnuCjo

TABLE 4. Experiment 3. Outcomes of the survey assessing the user experience. Results were collected for presence, co-presence and social presence along with their respective standard deviations, distinguishing whether the participant was performing the role of the Avatar or of the Avatarm.

	Avatarm	Avatar
Presence (P)	41.25% ± 30.65%	26.25% ± 28.65%
Co-Presence (CP)	48.13% ± 29.88%	32.5% ± 27.02%
Social Presence (SP)	2.68 ± 0.54	2.51 ± 0.47

and at the end of the experiment (i.e., once for each role), team members were asked to complete an online survey[6] to gather information on their experience. Each participant used the assigned anonymous identity (i.e., ''User1'', ''User2'', etc.). The survey consisted of 19 items extracted from [54], [55], and [56], of which four evaluating the sense of Presence (P), eight Co-Presence (CP), and seven Social Presence (SP). Some items from the original questionnaires were excluded as they were not consistent with our specific investigation. In accordance with [54] and [55], P and CP items were rated on a seven-point Likert scale, ranging from 1 to 7, and the respective scores were taken as the percentage of answers that had a score of '6' or '7'. Differently, following what proposed in [56], Social Presence items were rated on a five-point scale, ranging from 1 (*Strongly Disagreed*) to 5 (*Strongly Agreed*). Outcomes of the questionnaires are reported in Tab. 4.

V. DISCUSSION AND LIMITATIONS

The first experiment validates the performance and effectiveness of the algorithms for robot control and concealment. The gripper's positioning error is sufficiently small to perform daily activities effectively, such as pouring water or offering a glass. Although the results do not outperform existing state-of-the-art techniques, they demonstrate the robot control algorithm's suitability for integration into the entire system, despite the complexity and the possible delays caused by the interconnection of different parts. The concealment method, previously introduced in [5], is quantitatively assessed for the first time in this work. The obtained outcomes demonstrate high concealing quality, as our algorithm renders the robot nearly transparent in the augmented environment. These results are even more promising when considering that in the final configuration of the Physical Metaverse an avatar is superimposed onto the robot. This implies that those pixels of the robot remaining in the scene could be covered by the avatar, potentially enhancing the concealment even further.

The results of the second experiment confirm the crucial role of haptic feedback in accomplishing tasks like gentle manipulation of fragile objects. The haptic cue drastically reduced the occurrence of failures (i.e., the object falls or breaks) and minimized the contact force impulse during pick-and-place tasks. Moreover, these results are noteworthy as they demonstrate, for the first time, the successful simultaneous manipulation of objects in both virtual and real environments using the TouchDIVER. In our system, the force fed back to the user is not computed from the virtual interaction but is instead proportional to the measurements from gripper sensors. This highlights the successful integration of the haptic device into the overall framework of the Physical Metaverse.

Furthermore, the third experiment evaluated the effectiveness of the Avatarm in enhancing the sense of presence, co-presence, and social presence.

Since the Presence metric evaluates the feeling of being in a different environment without visual artifact, the optimal outcome should be having a low score for the user sharing the environment (the avatar) and higher score for the user interacting with the Avatarm. Indeed, the former should perceive the augmented environment as its surrounding augmented with a virtual entity (the Avatarm) and the latter should have the illusion of being in a different place. The obtained result of 26.25% ± 28.65% for the avatar role indicates that the user perceived its environment as barely altered (as they did not felt 'present' in a place other than their own), which is a promising result. On the other hand, a score of 41.25% ± 30.65% for the Avatarm indicates that the user is reasonably immersed in the remote environment. While there is still room for improvement, these results are encouraging and move in the right direction, namely, amplifying the sensation of being in a remote physical space thanks to Avatarm.

Similarly, the CP direction achieved better results in the Avatarm case, with a difference of 15.63%. In this case, the questions assessed to what extent users felt not alone in the environment. The stronger perception of the companion's presence experienced by the users in the role of the Avatarm may be attributed to the fact that participants controlling the Avatarm could see the actual companion through the video stream, while their own presence was represented by a digital avatar.

Lastly, users rated Social Presence as 2.68 ± 0.54 for the Avatarm and 2.51 ± 0.47 for the avatar. It is worth noting that, unlike previous cases, Social Presence results are reported as average score on a scale ranging from 0 to 5. Reminding that SP refers to the feeling of cooperating on a common task in a shared environment, the higher perception of SP experienced with the Avatarm can be attributed to the different roles users assumed in the task. When using the Avatarm, participants took control of the task, giving them both the responsibility and the ability to determine the experimental outcome. This likely fostered a greater sense of cooperation, as their actions were more dependent from those of their companion. On the contrary, the avatar was tasked with placing the initial eight cups, making their role somewhat independent from their companion's actions.

Overall, we observe that renewing object-based communication and enhancing interaction capabilities between remote companions positively impacts the user experience, as confirmed by the questionnaire results.

[6]https://forms.gle/KqAVaK6u76B1Bqe27

The conducted experiments also highlighted the current limitations of the resulting framework for the Physical Metaverse. The representation of the avatar in the shared environment appears unnatural and poorly blended with the background. Moreover, using two fixed cameras as viewing points for both the Avatarm and avatar users limits their ability to navigate and observe the surrounding scene. Finally, the current implementation of the concealing method imposes limitations on the full utilization of optical see-through devices for users co-located with the robotic arm (i.e., the avatar). Our algorithm requires a 2D panel to render the shared environment, effectively using the device as a screen rather than a transparent medium, even for viewing the surrounding environment. This limitation is less impactful for the Avatarm, as they need to view a different environment and cannot use a see-through device. However, rendering the scene in 2D diminishes the immersiveness of the experience for both users, and the lack of three-dimensionality remains a significant limitation.

VI. OPEN RESEARCH QUESTIONS AND FUTURE WORK

Despite the comprehensive experimental evaluation, several scientific questions remain unanswered. One of the open research questions arising from our work concerns the manipulation of objects that can change their state of matter. For instance, consider a bottle containing water that is being poured. In our work, we utilize digital twins that precisely replicate real-world objects. Consequently, it is crucial that the virtual environment also accurately reflects the same state changes. In the given example of water, as it is poured from the bottle, the digital twin in the virtual setting must dynamically adjust to depict the transition of water from one state to another. Furthermore, if the user who has been offered a glass of water drinks it, the water level must be updated accordingly in the virtual world and in a real-time manner.

In this regard, the level of realism for objects and avatars in the virtual scene represents a critical aspect that requires further investigation. On one side, ensuring high-quality textures that accurately resemble real shapes and colors of virtual entities would be desirable. On the other side, considering the famous effect of the Uncanny Valley [57], an interesting challenge is to evaluate the optimal level of realism for the Physical Metaverse that is preferred by users without causing discomfort. Understanding this balance is crucial for creating an engaging user experience.

Another open research question concerns the multi-user experience within the Physical Metaverse. Having more than one remote user joining the session necessitates the simultaneous use of multiple Avatarms, which introduces the complexity of coordinating multiple robots interacting together and potentially modifying the same remote environment. Understanding both the technological and emotional implications of this expanded experience is crucial for developing a more realistic Physical Metaverse where users can interact and socialize effectively without being limited to just two people.

TABLE 5. List of acronyms and abbreviations.

Acronym	Meaning
CAD	Computer Aided Design
CP	Co-Presence
DoF	Degree of Freedom
DR	Diminished Reality
HMD	Head Mounted Display
MAF	Moving Avarage Filter
MCP	MetaCarpoPhalangeal
NRMSE	Normalized Root Mean Square Error
P	Presence
RGBα	Red-Green-Blue-α
RMSE	Root Mean Square Error
ROS	Robotic Operating System
SP	Social Presence
UDP	User Data Protocol
XR	eXtended Reality

In terms of future work, the integration of additional types of cutaneous cues is certainly deserving of further development. Regarding feedback technology, the system is already equipped to facilitate this integration, as the TouchDIVER already incorporates three types of feedback. However, effort is needed to implement the automatic detection of these characteristics based on the grasped objects. This could involve utilizing sensors, such as a temperature sensor placed into the gripper, or employing vision-based algorithms for detecting textures, among other approaches. Addressing the current lack of avatar user's camera mobility could be faced by switching from a fixed camera to one attached to the HMD (or utilizing the built-in one when available). However, replacing the Avatarm viewpoint is more challenging, as the camera should track the head and torso movements of the remote interlocutor using a motorized support. Moreover, in the next version of Avatarm, we plan to integrate users' voices using microphones and speakers already built into the cameras (or into the HMD).

Lastly, as discussed in Section III-B2, future improvements involve developing an algorithm for user intention prediction and a switching control strategy for the robot end-effector. These advancements will enable the robot end-effector to move towards an object only when the user intends to grasp it, and then track the object's movement accordingly.

Addressing these limitations presents new challenges for future work, which are crucial for enhancing the user experience in the Physical Metaverse.

VII. CONCLUSION

This paper presented a novel framework for physical collaboration within the metaverse. This innovative concept is built

TABLE 6. List of symbols.

Sym	Definition	Sym	Definition
X	Length of images in pixels	m	Number of gripper jaws
Y	Height of images in pixels	$c_g(t)$	Gripper closure state
$F(t)$	3D matrix of image frames acquired by the digital camera	δc_g	Gripper closure step
$I(t)$	3D matrix of image frames acquired by the real camera	c_ν	Virtual hand closure state
$\wp(x,y)$	Pixel in coordinate (x,y) of image frame	μ	Hand to gripper scale factor
$M(t)$	Binary matrix, mask of image frame F(t)	$f_{r,i}$	The force applied by i-th jaw of the gripper
$H(t)$	3D matrix of resulting background of the Physiscal Metaverse	$\tau_{r,i}$	The torque applied by i-th jaw of the gripper
$\tilde{I}(t)$	3D matrix of resulting view of the Physiscal Metaverse user	λ_r	Generalized forces applied by gripper
$\varrho_i(t)$	User's palm pose acquired by tracking system	Λ_r	Maximum generalized forces applicable by gripper
$q_i(t)$	User's hand joint angles acquired by tracking system	$w_r(t)$	Wrench applied to the gravity center of a object
$\varrho_g(t)$	User's virtual palm pose rendered in XR scene	f_r	Force applied to the gravity center of a object
$q_g(t)$	User's virtual hand joint angles rendered in XR scene	τ_r	Torque applied to the gravity center of a object
V_j	Volume of intersection between j-th phalanx and an object	$G_r(t)$	Grasp matrix of gripper
t_γ	Time Instant of contact between virtual hand and digital objects	$P_{r,j}$	Rotation from the j-th contact to inertial frame
γ	Stop signal to avoid virtual hand unrealistic indentation	$P_{r,j}$	Displacement from j-th contact point to inertial frame
$\mathbb{W}$	Real world reference frame	$S(\cdot)$	Cross product matrix of a generic vector
$\mathbb{V}$	Digital world reference frame	$\mathcal{N}(\cdot)$	Null space of a generic matrix
e	Robot end-effector	ξ_r	Squeezing force applied by gripper jaw
o	Digital twin of an object	$\sigma_r(t)$	Magnitude of squeezing forces applied by whole gripper
${}^i p_j(t)$	Position of j in reference frame i	K_r	Selection matrix of forces to render on virtual object
δt	MAF time-window	η	Gripper to haptic interface scale factor
${}^i\theta_j(t)$	Orientation of j in reference frame i	$\sigma_v(t)$	Magnitude of squeezing forces on virtual object
${}^i\varrho_j$	Pose (position and orientation) of j in reference frame i	$\xi_h(t)$	Squeezing force to render by haptic interface
${}^i T_j$	Homogeneous transformation expressing j pose in reference frame i	K_ν	Selection matrix of forces to render with haptic interface
${}^i A_j$	Auxiliary transformation expressing j pose in reference frame i	$F_{r,j}$	Robot force limit
${}^i R_j(t)$	Rotation matrix expressing j orientation in reference frame i	$T_{r,j}$	Robot torque limit
${}^i\nu_j(t)$	Generalized velocity of j expressed in reference frame i	$\Xi_{r,f,j}$	Haptic interface force limit
${}^i\dot{p}_j(t)$	Linear velocity of j expressed in reference frame i	$\Xi_{r,\tau,j}$	Haptic interface torque limit
${}^i\omega_j(t)$	Angular velocity of j expressed in reference frame i	$\chi(t)$	Manipulability measure
$r_{\dot{p}}(t)$	Scale factor to accomplish the ISO security standard on $\dot{w}_e^p$	h	Haptic vibrotactile signal
$\psi(t)$	Scale factor to accomplish the ISO security standard on $\dot{w}_e^\nu$	$[a,b]$	Haptic signal parameters
J	Robot Jacobian matrix	$\mathcal{M}_\%$	Goodness measure of the diminishing algorithm
n	Number of interactive finger (excluding thumb)	N	Number of video frames in a video recording
$\tilde{h}$	Max digital value of haptic interface	$\tilde{\chi}$	Manipulability lower bound

upon the Avatarm, an advanced avatar with the capability to interact with physical objects. This introduces a new form of extended remote physical environment shared among multiple users, which we term the 'Physical Metaverse'.

This achievement is made possible thanks to i) a robotic manipulator that grasps, moves, and places the objects in a remote physical environment following the trajectory traced by the object's digital twin, ii) software that diminishes the video stream of the robot, and iii) haptic interfaces for rendering the forces applied at the end-effector and making the user who controls the Avatarm aware of the state of the hidden robot. In this way, the Physical Metaverse goes beyond the digital boundaries of the metaverse transforming the movement of digital objects into the motion of their tangible counterpart.

Together with the characterization of each component enabling the Avatarm, we described its implementation and we provided the results of an experimental validation aimed at testing the overall performance. These demonstrated the efficacy of the proposed framework, and shed light on some areas for improvement, as highlighted in Section VI. However, by continuing to work to overcome these limitations, the Avatarm may not only shape a new kind of XR experience but also redefine how we interact with and perceive our surroundings when immersed in the metaverse, making it tangible and physical.

APPENDIX
ACRONYMS AND SYMBOLS

See Table 6.

REFERENCES

[1] A. Davis, J. Murphy, D. Owens, D. Khazanchi, and I. Zigurs, "Avatars, people, and virtual worlds: Foundations for research in metaverses," *J. Assoc. Inf. Syst.*, vol. 10, no. 2, pp. 90–117, Feb. 2009.

[2] R. V. Kozinets, "Immersive netnography: A novel method for service experience research in virtual reality, augmented reality and metaverse contexts," *J. Service Manage.*, vol. 34, no. 1, pp. 100–125, Jan. 2023.

[3] G. Riva, *Interreality: A New Paradigm for E-Health*. Amsterdam, The Netherlands: IOS Press, 2009.

[4] G. Riva and B. K. Wiederhold, "What the metaverse is (really) and why we need to know about it," *Cyberpsychol., Behav., Social Netw.*, vol. 25, no. 6, pp. 355–359, 2022.

[5] A. Villani, G. Cortigiani, B. Brogi, N. D'Aurizio, T. Lisini Baldi, and D. Prattichizzo, "Avatarm: An avatar with manipulation capabilities for the physical metaverse," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11626–11632.

[6] A. F. Siu, S. Yuan, H. Pham, E. Gonzalez, L. H. Kim, M. Le Goc, and S. Follmer, "Investigating tangible collaboration for design towards augmented physical telepresence," in *Design Thinking Research: Making Distinctions: Collaboration Versus Cooperation*. Cham, Switzerland: Springer, 2018, pp. 131–145.

[7] T. Aoki, T. Kuriyama, K. Asano, T. Kawase, I. Matumura, T. Matsushita, Y. Iio, H. Mitake, T. Toyama, S. Hasegawa, R. Ayukawa, H. Ichikawa, and M. Sato, "Kobito: Virtual brownies," in *Proc. ACM SIGGRAPH Emerg. Technol.*, 2005, p. 11.

[8] M. Lee, N. Norouzi, G. Bruder, P. J. Wisniewski, and G. F. Welch, "The physical-virtual table: Exploring the effects of a virtual human's physical influence on social interaction," in *Proc. 24th ACM Symp. Virtual Reality Softw. Technol.*, Nov. 2018, pp. 1–11.

[9] R. Suzuki, A. Karim, T. Xia, H. Hedayati, and N. Marquardt, "Augmented reality and robotics: A survey and taxonomy for AR-enhanced human–robot interaction and robotic interfaces," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2022, pp. 1–33.

[10] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-hand behavior in human–robot shared manipulation," in *Proc. 13th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2018, pp. 4–13.

[11] Y. F. Cheng, H. Yin, Y. Yan, J. Gugenheimer, and D. Lindlbauer, "Towards understanding diminished reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2022, pp. 1–16.

[12] A. V. Taylor, A. Matsumoto, E. J. Carter, A. Plopski, and H. Admoni, "Diminished reality for close quarters robotic telemanipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 11531–11538.

[13] M. Schwarz, C. Lenz, A. Rochow, M. Schreiber, and S. Behnke, "NimbRo avatar: Interactive immersive telepresence with force-feedback telemanipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 5312–5319.

[14] B. Jones, Y. Zhang, P. N. Wong, and S. Rintel, "Belonging there: VROOM-ing into the uncanny valley of XR telepresence," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, pp. 1–31, Apr. 2021.

[15] D.-I.-D. Han, Y. Bergs, and N. Moorhouse, "Virtual reality consumer experience escapes: Preparing for the metaverse," *Virtual Reality*, vol. 26, no. 4, pp. 1443–1458, Dec. 2022.

[16] M. Lombard and T. Ditton, "At the heart of it all: The concept of presence," *J. Comput.-Mediated Commun.*, vol. 3, no. 2, Jun. 2006, Art. no. JCMC321.

[17] H. T. Regenbrecht, T. W. Schubert, and F. Friedmann, "Measuring the sense of presence and its relations to fear of heights in virtual environments," *Int. J. Hum.-Comput. Interact.*, vol. 10, no. 3, pp. 233–249, Sep. 1998.

[18] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, "Affective interactions using virtual reality: The link between presence and emotions," *CyberPsychol. Behav.*, vol. 10, no. 1, pp. 45–56, Feb. 2007.

[19] M. Slater, D.-P. Pertaub, and A. Steed, "Public speaking in virtual reality: Facing an audience of avatars," *IEEE Comput. Graph. Appl.*, vol. 19, no. 2, pp. 6–9, Mar. 1999.

[20] M. Stylianos, "Metaverse," *Encyclopedia*, vol. 2, no. 1, pp. 486–497, 2022.

[21] C. Youngblut, "Experience of presence in virtual environments," Inst. Defense Analyses, Alexandria, VA, USA, Tech. Rep. ADA427495, 2003. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA427495

[22] F. Biocca, "The Cyborg's dilemma: Progressive embodiment in virtual environments," *J. Comput.-Mediated Commun.*, vol. 3, no. 2, Jun. 2006, Art. no. JCMC324.

[23] B. E. Mennecke, J. L. Triplett, L. M. Hassall, Z. J. Conde, and R. Heer, "An examination of a theory of embodied social presence in virtual worlds," *Decis. Sci.*, vol. 42, no. 2, pp. 413–450, 2011.

[24] C. Eckert and J.-F. Boujut, "The role of objects in design co-operation: Communication through physical or virtual objects," *Comput. Supported Cooperat. Work*, vol. 12, no. 2, pp. 145–151, Jun. 2003.

[25] M. Brereton and B. McGarry, "An observational study of how objects support engineering design thinking and communication: Implications for the design of tangible media," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2000, pp. 217–224.

[26] D. Nicolini, J. Mengis, and J. Swan, "Understanding the role of objects in cross-disciplinary collaboration," *Org. Sci.*, vol. 23, no. 3, pp. 612–629, Jun. 2012.

[27] C. Bueger and J. Stockbruegger, "Actor-network theory: Objects and actants, networks and narratives," in *Technology and World Politics*. Evanston, IL, USA: Routledge, 2017, pp. 42–59.

[28] C. Pacchierotti and D. Prattichizzo, "Cutaneous/tactile haptic feedback in robotic teleoperation: Motivation, survey, and perspectives," *IEEE Trans. Robot.*, vol. 40, pp. 978–998, 2024.

[29] C. Bermejo and P. Hui, "A survey on haptic technologies for mobile augmented reality," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–35, Dec. 2022.

[30] R. M. Pierce, E. A. Fedalei, and K. J. Kuchenbecker, "A wearable device for controlling a robot gripper with fingertip contact, pressure, vibrotactile, and grip force feedback," in *Proc. IEEE Haptics Symp. (HAPTICS)*, Feb. 2014, pp. 19–25.

[31] *Weart TouchDIVER*. Accessed: Mar. 2024. [Online]. Available: https://www.weart.it/touchdiver/

[32] M. Ferro, C. Pacchierotti, S. Rossi, and M. Vendittelli, "Deconstructing haptic feedback information in robot-assisted needle insertion in soft tissues," *IEEE Trans. Haptics*, vol. 16, no. 4, pp. 536–542, Oct./Dec. 2023, doi: 10.1109/TOH.2023.3271224.

[33] N. Kosanovic, J. C. Vaz, and P. Y. Oh, "Biomimetic real-time multimodal tactile perception and haptics for telepresence humanoids," in *Proc. 21st Int. Conf. Adv. Robot. (ICAR)*, Dec. 2023, pp. 613–620.

[34] A. Pedersen, M. Jørgensen, H. Isaksen, and M. Riedel, "Playing the virtual glass armonica," in *Proc. 19th Sound Music Comput. Conf. (SMC)*, 2022, pp. 672–673.

[35] Ö. E. Dural, A. A. Shahid, G. Gioioso, D. Prattichizzo, F. Braghin, and L. Roveda, "Evaluation of tactile feedback for teleoperated glove-based interaction tasks," in *Proc. Int. Workshop Hum.-Friendly Robot.* Cham, Switzerland: Springer, 2023, pp. 79–93.

[36] S. Kumar, C. Savur, and F. Sahin, "Survey of human–robot collaboration in industrial settings: Awareness, intelligence, and compliance," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 280–297, Jan. 2021.

[37] A. Casalino, C. Messeri, M. Pozzi, A. M. Zanchettin, P. Rocco, and D. Prattichizzo, "Operator awareness in human–robot collaboration through wearable vibrotactile feedback," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4289–4296, Oct. 2018.

[38] T. Richter-Trummer, D. Kalkofen, J. Park, and D. Schmalstieg, "Instant mixed reality lighting from casual scanning," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Sep. 2016, pp. 27–36.

[39] N. Kulkarni, "Color thresholding method for image segmentation of natural images," *Int. J. Image, Graph. Signal Process.*, vol. 4, no. 1, pp. 28–34, Feb. 2012.

[40] J. Mezirow, "Perspective transformation," *Adult Educ.*, vol. 28, no. 2, pp. 100–110, 1978.

[41] A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen, "Automatic grasp planning using shape primitives," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 2, Sep. 2003, pp. 1824–1829.

[42] Y. Lin and Y. Sun, "Robot grasp planning based on demonstrated grasp strategies," *Int. J. Robot. Res.*, vol. 34, no. 1, pp. 26–42, Jan. 2015.

[43] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4304–4311.

[44] G. Saponaro, G. Salvi, and A. Bernardino, "Robot anticipation of human intentions through continuous gesture recognition," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2013, pp. 218–225.
[45] W. Miao, G. Li, G. Jiang, Y. Fang, Z. Ju, and H. Liu, "Optimal grasp planning of multi-fingered robotic hands: A review," *Appl. Comput. Math.*, vol. 14, no. 3, pp. 238–247, 2015.
[46] A. Gasparetto, P. Boscariol, A. Lanzutti, and R. Vidoni, "Path planning and trajectory planning algorithms: A general overview," *Motion Oper. Planning Robot. Syst.*, vol. 29, pp. 3–27, Jan. 2015.
[47] *Robots for Industrial Environments—Safety Requirements—Part 1: Robot*, Standard ISO10218-1:2006, ISO, 2006.
[48] J. Lee and T. L. Kunii, "Model-based analysis of hand posture," *IEEE Comput. Graph. Appl.*, vol. 15, no. 5, pp. 77–86, Sep. 1995.
[49] T. Lisini Baldi, S. Scheggi, L. Meli, M. Mohammadi, and D. Prattichizzo, "GESTO: A glove for enhanced sensing and touching based on inertial and magnetic sensors for hand tracking and cutaneous feedback," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 6, pp. 1066–1076, Dec. 2017.
[50] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer Handbook of Robotics*. Berlin, Germany: Springer, 2016, pp. 955–988.
[51] G. Salvietti, L. Meli, G. Gioioso, M. Malvezzi, and D. Prattichizzo, "Multicontact bilateral telemanipulation with kinematic asymmetries," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 1, pp. 445–456, Feb. 2017.
[52] C. Gaudeni, T. Lisini Baldi, G. M. Achilli, M. Mandalà, and D. Prattichizzo, "Instrumenting hand-held surgical drills with a pneumatic sensing cover for haptic feedback," in *Proc. Int. Conf. Hum. Haptic Sens. Touch Enabled Comput. Appl.* Cham, Switzerland: Springer, 2020, pp. 398–406.
[53] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set," *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 36–52, Sep. 2015.
[54] M. Usoh, E. Catena, S. Arman, and M. Slater, "Using presence questionnaires in reality," *Presence*, vol. 9, no. 5, pp. 497–503, Oct. 2000.
[55] C. Basdogan, C.-H. Ho, M. A. Srinivasan, and M. Slater, "An experimental study on the role of touch in shared virtual environments," *ACM Trans. Comput.-Hum. Interact.*, vol. 7, no. 4, pp. 443–460, Dec. 2000.
[56] C. N. Gunawardena and F. J. Zittle, "Social presence as a predictor of satisfaction within a computer-mediated conferencing environment," *Amer. J. Distance Educ.*, vol. 11, no. 3, pp. 8–26, Jan. 1997.
[57] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012.

BERNARDO BROGI (Student Member, IEEE) received the B.Sc. degree in computer engineering and the M.Sc. degree (cum laude) in artificial intelligence and automation engineering from the University of Siena, Siena, Italy, in 2020 and 2022, respectively. He is currently pursuing the Ph.D. degree in robotics and intelligent machines, administrated by the University of Genoa and hosted at the University of Siena. His research interests include robotics and extended reality, with a focus on human–robot collaboration for healthcare and wellness of persons.

GIOVANNI CORTIGIANI (Student Member, IEEE) received the B.Sc. degree (cum laude) in computer engineering and the M.Sc. degree (cum laude) in artificial intelligence and automation engineering from the University of Siena, Siena, Italy, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree with the Department of Information Engineering and Mathematics. His research interests include robotics and virtual reality, with a focus on human–robot collaboration, manipulation, and grasping.

ALBERTO VILLANI (Student Member, IEEE) received the B.Sc. and M.Sc. degrees (cum laude) in automation engineering from the University of Naples, Italy, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree in robotics and automation with the Department of Information Engineering and Mathematics, University of Siena. His research interests include robotics and haptics, with a focus on human- and cell-centered robotics.

NICOLE D'AURIZIO (Member, IEEE) received the M.Sc. degree (cum laude) in computer and automation engineering from the University of Siena, in 2019, and the Ph.D. degree in automatic control and robotics from the Department of Information Engineering and Mathematics, University of Siena, in 2023. She was a Ph.D. Fellow with the Department of Advanced Robotics, Istituto Italiano di Tecnologia (IIT), from 2019 to 2022. She is currently an Assistant Professor with the Department of Information Engineering and Mathematics, University of Siena, and a Research Affiliate with IIT. Her research interests include haptics and robotics, with a focus on wearable and affective haptics and human-centered robotics.

DOMENICO PRATTICHIZZO received the M.S. degree in electronics engineering and the Ph.D. degree in robotics and automation from the University of Pisa, Pisa, Italy, in 1991 and 1995, respectively. In 1994, he was a Visiting Scientist with the MIT AI Laboratory. Since 2009, he has been a Scientific Consultant with the Istituto Italiano di Tecnologia, Genoa, Italy. He is currently a Professor of robotics with the University of Siena, Siena, Italy. He has authored more than 200 articles in his research fields. His main research interests include haptics, grasping, visual servoing, and mobile robotics.

TOMMASO LISINI BALDI (Member, IEEE) received the M.Sc. degree (cum laude) in computer and automation engineering and the Ph.D. degree in robotic and automation from the Department of Information Engineering and Mathematics, University of Siena, Siena, Italy, in 2014 and 2018, respectively. He was a Ph.D. Fellow with the Department of Advanced Robotics, Istituto Italiano di Tecnologia, from 2014 to 2017. He is currently an Assistant Professor with the University of Siena and a Research Affiliate with the Istituto Italiano di Tecnologia. His research interests include robotics and haptics, with a focus on human–robot collaboration, haptics feedback, and motion tracking with inertial sensors.