

# NimbRo wins ANA Avatar XPRIZE Immersive Telepresence Competition: Human-Centric Evaluation and Lessons Learned

Christian Lenz<sup>1\*†</sup>, Max Schwarz<sup>1†</sup>, Andre Rochow<sup>1</sup>, Bastian Pätzold<sup>1</sup>,  
Raphael Memmesheimer<sup>1</sup>, Michael Schreiber<sup>1</sup>, Sven Behnke<sup>1</sup>

<sup>1</sup>Autonomous Intelligent Systems, University of Bonn, Germany.

\*Corresponding author(s). E-mail(s): [lenz@ais.uni-bonn.de](mailto:lenz@ais.uni-bonn.de);

Contributing authors: [schwarz@ais.uni-bonn.de](mailto:schwarz@ais.uni-bonn.de); [rochow@ais.uni-bonn.de](mailto:rochow@ais.uni-bonn.de);  
[paetzold@ais.uni-bonn.de](mailto:paetzold@ais.uni-bonn.de); [memmesheimer@ais.uni-bonn.de](mailto:memmesheimer@ais.uni-bonn.de); [schreiber@ais.uni-bonn.de](mailto:schreiber@ais.uni-bonn.de);  
[behnke@cs.uni-bonn.de](mailto:behnke@cs.uni-bonn.de);

†These authors contributed equally to this work.

## Abstract

Robotic avatar systems can enable immersive telepresence with locomotion, manipulation, and communication capabilities. We present such an avatar system, based on the key components of immersive 3D visualization and transparent force-feedback telemanipulation. Our avatar robot features an anthropomorphic upper body with dexterous hands. The remote human operator drives the arms and fingers through an exoskeleton-based operator station, which provides force feedback both at the wrist and for each finger. The robot torso is mounted on a holonomic base, providing omnidirectional locomotion on flat floors, controlled using a 3D rudder device. Finally, the robot features a 6D movable head with stereo cameras, which stream images to a VR display worn by the operator. Movement latency is hidden using spherical rendering. The head also carries a telepresence screen displaying an animated image of the operator's face, enabling direct interaction with remote persons. Our system won the \$10M ANA Avatar XPRIZE competition, which challenged teams to develop intuitive and immersive avatar systems that could be operated by briefly trained judges. We analyze our successful participation in the semifinals and finals and provide insight into our operator training and lessons learned. In addition, we evaluate our system in a user study that demonstrates its intuitive and easy usability.

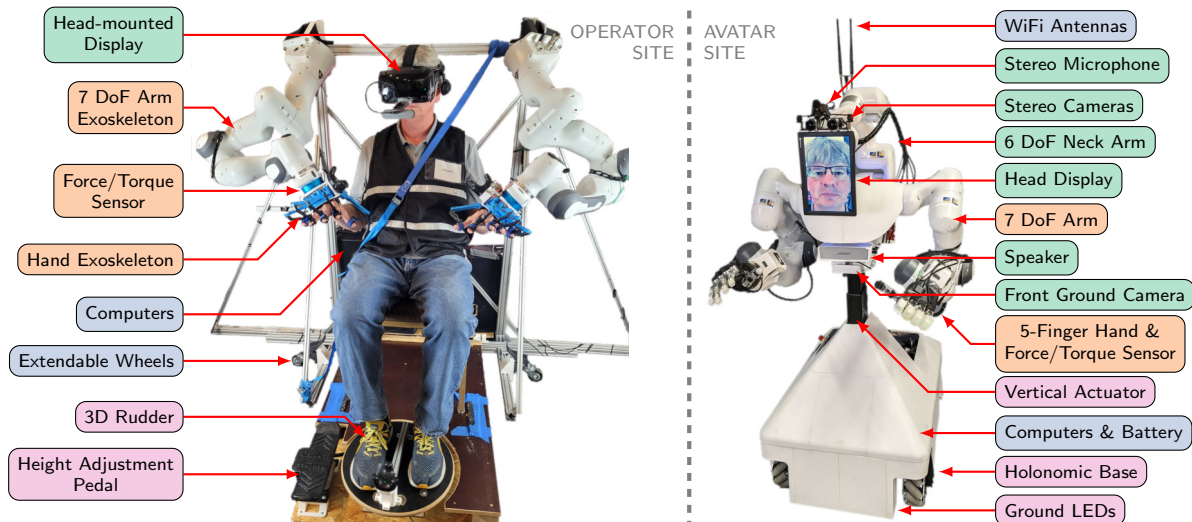
**Keywords:** Telepresence, Avatar, Dual-Arm Teleoperation, Face Animation, Competitions

## 1 Introduction

Robotic avatar systems combine high-quality telecommunication with intuitive robotic teleoperation, creating true telepresence. These systems allow full immersion into a remote space while also *embodying* the operator in a robotic system, giving them the ability to navigate in the remote

environment, to manipulate objects there, and to interact with remote persons in a multimodal way that includes direct physical contact.

Through such an immersive telepresence system, humans can perform tasks in remote environments which are currently beyond the capabilities of autonomous perception, planning, and control methods—the human intellect is still unmatched



**Fig. 1** NimbRo avatar system consisting of the operator station (left) and the avatar robot (right). Colors correspond to categories (Audiovisual telepresence, Telemanipulation, Locomotion, System / Misc).

in its ability to perceive, plan, and react to unforeseen situations. Avatar systems allow humans to work in remote environments without having to travel or expose themselves to potential dangers, such as in disaster response.

The ANA Avatar XPRIZE competition<sup>1</sup> challenged the robotics community to advance the state of the art in immersive telepresence systems. Equipped with a record \$10M prize purse, the competition required teams to build *intuitive* and *robust* robotic avatar systems that allow a human operator to be present in a remote space. The tasks to be solved included social interaction and communication, but also locomotion and complex manipulation. Critically, the systems were to be used and evaluated by operator and recipient judges. In contrast to previous teleoperation competitions such as the DARPA Robotics Challenge [1], operators could be trained only for a short time to use the developed avatar systems.

Our NimbRo Avatar system (Fig. 1) won the ANA Avatar XPRIZE Finals in November 2022<sup>2</sup>. This article aims to provide a comprehensive overview of our system by summarizing previous publications [2–7]. It also extends these previous works by focusing on the human interaction, both between operator and machine as well as operator

and human recipients interacting with the avatar. Our contributions include

- an avatar robot with humanoid upper body and the accompanying operator station,
- bidirectional audiovisual telepresence including latency-free rendering during head movement as well as operator face animation,
- transparent dual-arm telemanipulation with force feedback on arm and finger level,
- roughness sensing and haptic display,
- safety, monitoring, and robustness modules,
- a detailed analysis of our competition performance as well as a user study, and
- a discussion of lessons learned.

## 2 Related Work

Telemanipulation robots are complex systems consisting of many components, which have been investigated both individually as well as on a systems level.

### *Telemanipulation Systems*

The DARPA Robotics Challenge (DRC) 2015 [1] resulted in the development of several mobile telemanipulation robots, such as DRC-HUBO [8], CHIMP [9], RoboSimian [10], and our own entry Momaro [11]. All these systems demonstrated impressive locomotion and manipulation capabilities under teleoperation, even with severely

<sup>1</sup><https://www.xprize.org/prizes/avatar>

<sup>2</sup><https://www.youtube.com/watch?v=EmESa2Olq4c>

constrained communication. The DRC placed no emphasis on intuitiveness of the teleoperation controls or on immersion of the operators, though. To our knowledge, our team was the only one using a VR head-mounted display (HMD) and 6D magnetic trackers to perceive the environment in 3D and to control the robot arms [12]—all other teams relying entirely on 2D monitors and traditional input devices to control their robots. All teams, including ours, required highly trained operators familiar with the custom-designed operator interfaces. Furthermore, since the DRC was geared towards disaster response, the robots did not feature any communication capabilities for interacting with remote humans.

In our subsequent work [13], we developed the ideas embodied in the Momaro system further. The resulting Centauro robot is a torque-controlled platform capable of locomotion and dexterous manipulation in rough terrain. It is controlled by a human sitting in a dedicated operator station, equipped with an upper-body exoskeleton providing force feedback and a VR HMD. Still, Centauro is focused on disaster response and does not have any communication facilities.

Recently, there has been explosive growth in teleoperated humanoid robots. Darvish *et al.* [14] provide a comprehensive overview comparing fully-humanoid (i.e. walking) robotic teleoperation systems. Walking humanoid robots can overcome obstacles, rough terrain, and stairs, but raise more complex challenges regarding balance and whole-body control. As the XPRIZE competition as well as many human-made environments feature flat surfaces on which wheels are more efficient and stable, our system features a wheeled omnidirectional base.

Schmaus *et al.* [15] discuss the results of the METRON SUPVIS Justin space-robotics experiment, where an astronaut on the international space station controlled the Justin robot on Earth, simulating an orbital robotics mission. Instead of opting for full immersion and direct control, the authors relied on a 2D tablet display and higher levels of autonomy, allowing the astronaut to trigger autonomous task skills. A similar approach was developed for our domestic service robot Cosero [16].

In contrast to the discussed prior works, our avatar system is specifically designed to operate in human workspaces and to interact with

humans. To the best of our knowledge, there were no integrated robots designed for this purpose prior to the ANA Avatar XPRIZE, which initiated development of such systems [17–23]. Notably, Luo *et al.* [17] describe the approach by the third-placed Team Northeastern, who developed hydraulic grippers with high-fidelity force feedback. Similar to our system, the team incorporated two Franka Emika Panda arms for bimanual manipulation on top of a omnidirectional base. Team AVATRINA, described by Marques *et al.* [18], reached fourth place, again with a bimanual manipulation system based on Franka Emika Panda arms. In contrast to other top-placed teams, our system featured increased immersion and intuitiveness through free 6D head movement, a photorealistic face animation system for communication with recipients, and haptic feedback from multimodal sensors. In addition, our team focused on system robustness, rigorous testing, and training of the crew that later trained operators to use the avatar system.

### 3D VR Televisualization

Live capture and visualization of the remote scene is typically done using data from RGB or RGB-D cameras. There are many examples of static and movable stereo cameras on robots, which are directly visualized in a head-mounted display [24–26]. However, these approaches are limited by either a fixed viewpoint or considerable camera movement latency, potentially creating motion sickness. In contrast, our system hides latencies by correcting viewpoint changes through spherical rendering [3].

RGB-D sensors allow rendering from free viewpoints [27, 28], removing head movement latency. However, these sensors produce depth images with missing measurements, which can be difficult to visualize in a convincing way. Reconstruction-based approaches [12, 13, 29] address this issue by aggregating 3D measurements over time and building dense representations, which can be viewed without movement latency. They still, however, struggle with many reflective and transparent materials, because the depth sensors cannot measure them. An additional drawback is that reconstruction-based approaches usually cannot deal with dynamic scenes—which is an issue when interacting with the environment and human

recipients. In contrast, our method always displays a live stereo RGB stream, which has no difficulties with materials or dynamic scenes.

### Force Feedback

Teleoperation systems use typically stationary devices to display any force feedback captured by the remote robot to the human operator [13, 30, 31]. In contrast, wearable haptic devices [32] are usually more lightweight and do not limit the operator’s workspace. However, they cannot display absolute forces to the operator.

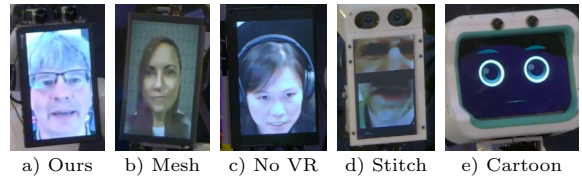
Much recent and ongoing research focuses on stable teleoperation systems in time-delayed scenarios [33, 34]. Large time delays for teleoperation in earth-space scenarios are investigated [35, 36]. In our application, we assume smaller distances between the operator station and the avatar robot. Thus, our force feedback controller does not need to handle such high latencies.

### Locomotion Control

Locomotion control is a key aspect of avatar robot systems. Directly related are locomotion interfaces for virtual reality control. Interesting hands-free locomotion control can be achieved by e.g. treadmills [37], circular moving tiles [38], or walking pads [39]. These approaches tend to be exhausting for the operator and are not suitable for long-term operation. Further, in our setup they might transfer unintended motion to the robot’s arms. More relaxing for the operators are seated locomotion controllers. Ohshima *et al.* [40] present a device that integrates a pressure sensor in a seat cushion. To detect lifting of the operator’s leg which is then translated to motion commands. Some interfaces are feet-controlled. Carmichael *et al.* [41] present a device that is attached to the operator’s feet and is able to detect the operators foot movement. Otaran & Farkhatdinov [42] introduced a feet controller for linear movement by recognizing steps on the platform. Interestingly, they integrate haptic feedback e.g. for the terrain. Other approaches also consider leaning as input [43].

### Facial Animation

Visualizing facial expressions of people wearing VR HMDs is a well-known task enabling remote social interaction. Often, eye tracking cameras capture eye poses and expressions such as frowns,



**Fig. 2** Types of facial animation at ANA Avatar XPRIZE finals. Examples from teams: a) NimbRo, b) Pollen Robotics, c) Northeastern [49], d) AVATRINA [18], e) UNIST.

while a standard camera captures the unobscured lower part of the face[44–46]. A special requirement of the ANA Avatar XPRIZE competition was that the method had to be quickly adaptable to a new operator, as only 45 min of setup time was allowed.

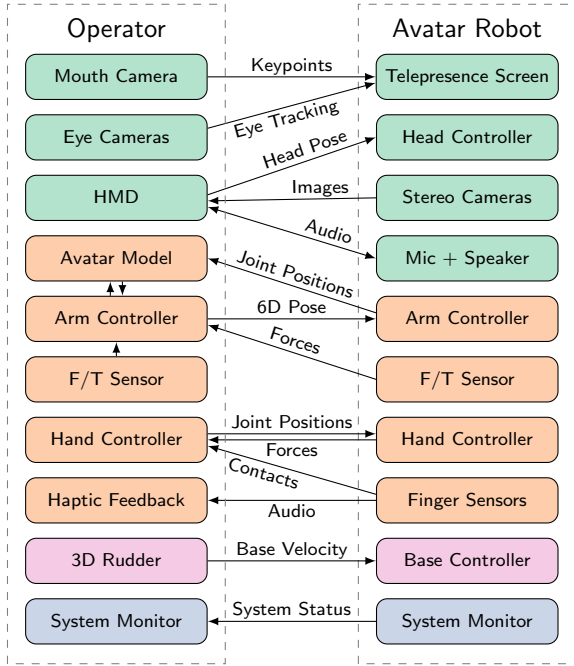
A first category of HMD facial animation methods is based on explicit 3D representations. Olszewski *et al.* [47] train a neural regressor to output blend shape weights, which deform a face mesh. On the other hand, Codec Avatars [44–46, 48] are an implicit model trained on many images of the operator.

All the mentioned methods require either extensive manual work (3D modeling), complicated capture setups (3D reconstruction), or long training times, all of which were infeasible in the avatar competition. In contrast, our 2D approach is based on taking a short video of the operator and does not require any on-site or operator-specific training.

From ANA Avatar XPRIZE finals video footage we recognize five categories of face animation techniques used by participants (see Fig. 2). Our team was the only one to produce a photorealistic animated face image.

## 3 NimbRo Avatar System

The NimbRo avatar system consists of the operator station and the avatar robot, which allows a human operator to feel present and interact in a remote location (Figs. 1 and 3). The operator station includes two arm and hand exoskeletons for telemanipulation with force and haptic feedback, a Head Mounted Display (HMD) that transmits video and audio for immersive telepresence, and two foot devices for locomotion control and avatar height adjustment. All components are connected via a standard PC (AMD Ryzen 9 5950X @



**Fig. 3** Information flow between the system components. Same coloring as in Fig. 1.

3.4 GHz, NVIDIA RTX A6000) and communicate via the Robot Operating System (ROS) [50] framework. The entire operator station can be moved by extending four wheels and can be temporarily powered by an EcoFlow portable power station for about 90 min.

The avatar robot has an anthropomorphic upper body with two arms ending in five-fingered hands, a head carrying a pair of stereo cameras, a stereo microphone, and a telepresence screen displaying the operator’s face. The upper body is attached to the mobile omnidirectional base through a height-adjustable spine. The robot footprint is  $90 \times 62$  cm and its height is between 128 and 182 cm. A shoulder width of 78 cm allows navigation through narrow passages such as standard doors. The total weight including battery and all onboard computing is approx. 140 kg. The robot is powered by a RELiON InSight 48 V 30 Ah battery, which allows approx. 2 hours of operation. Its base contains three arm controllers, power supplies, and a PC (Intel i9 12900K @ 5.2 GHz, NVIDIA RTX 3070) running a dedicated ROS instance. We describe the individual system components with more detail in the following sections.

## 4 Audiovisual Telepresence

Convincing telepresence requires both seeing and hearing as well as being seen and being heard. Conventional video conference systems cover this functionality already quite well and it is important for acceptance that robotic avatar systems do not fall behind these. We will now detail parts of our system that achieve audiovisual telepresence going beyond video conferencing while coping with the additional challenges imposed by a robotic avatar system.

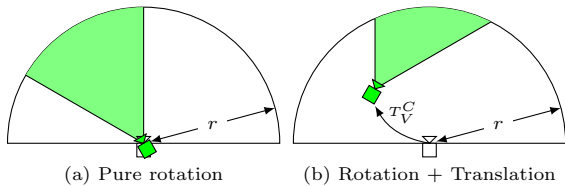
### 4.1 Robot Cameras & VR Display

The robot head is equipped with two wide-angle Basler a2A3840-45ucBAS cameras with an optical frame distance of 64 mm—matching the average human pupillary distance [51]. We crop the stereo video stream to a resolution of  $2 \times 2472 \times 2178$  @ 46 Hz, which gives a horizontal and vertical FoV of approximately  $160^\circ$ . The stream is then compressed with very low latency on the robot and transmitted over WiFi (see Sec. 7). The operator wears a Valve Index VR head-mounted display, which renders a view onto the remote scene. Using a VR HMD leads to full operator immersion. The cameras are intrinsically and extrinsically calibrated to be able to render the operator view correctly (see Sec. 5.5).

### 4.2 6D Movable Head & Spherical Rendering

In contrast to all other teams in the ANA Avatar XPRIZE, our robot has a separate 6 Degree of Freedom (DoF) arm (UFactory xArm6) which carries the head. This enables the robot to mirror *all* head movements made by the operator, in contrast to the common pan/tilt neck joints which only allow 2 DoF rotation. As a consequence, the operator can look *around* objects simply by translating their head, as they would if they were present in the remote scene. They also can choose viewpoints for manipulation that minimize occlusion. Finally, recipients interacting with the avatar frequently note that the full head movements contribute to the liveliness and identification of the avatar with the operator.

Allowing head movements, comes with a problem, though. Since there is considerable movement latency (roughly 200 ms) introduced by the



**Fig. 4** Spherical rendering example in 2D. We show only one camera  $C$  of the stereo pair, the other is processed analogously. The robot camera is shown in white with its very wide FoV. The corresponding VR camera  $V$ , which renders the view displayed to the operator, is shown in green. The camera image is projected onto the sphere with radius  $r$ , and then back into the VR camera. Pure rotations (a) result in no distortion while translations (b) will distort the image if the objects are not exactly at the assumed depth. Figure adapted from Schwarz & Behnke [3].

**Table 1** Head movement ablations.

Visual mode	Success	Completion time [s]	
		Mean	StdDev
a) Full 6D	7/7	71.0	50.6
b) 3D orientation	7/7	111.7	87.5
c) Fixed perspective	6/7	158.3	39.6

Source: Schwarz & Behnke [3].

masses, friction, and motor velocity constraints, directly rendering the video stream on the HMD results in operator motion sickness. Instead, we use a technique which renders each frame in a sphere centered on the capture location (see Fig. 4). The operator moves freely and with low latency inside this sphere. While head translations induce transient distortions (see Fig. 4), head rotations, which are most affected by latency due to the large lever effect for distant objects, can be perfectly handled. For more details about this technique, we refer to Schwarz & Behnke [3].

We evaluated the 6D neck joint against ablations with only 3D rotation or a fixed perspective in a small user study, which showed advantages for the full 6D mode (see Tab. 1). Additionally, operators reported that the freedom of movement in 6D was helpful for an insertion task, where viewing from the side was beneficial [3]. The benefits of stereoscopic vision itself for teleoperation have been shown before, e.g. by Triantafyllidis *et al.* [53].

### 4.3 Facial Animation

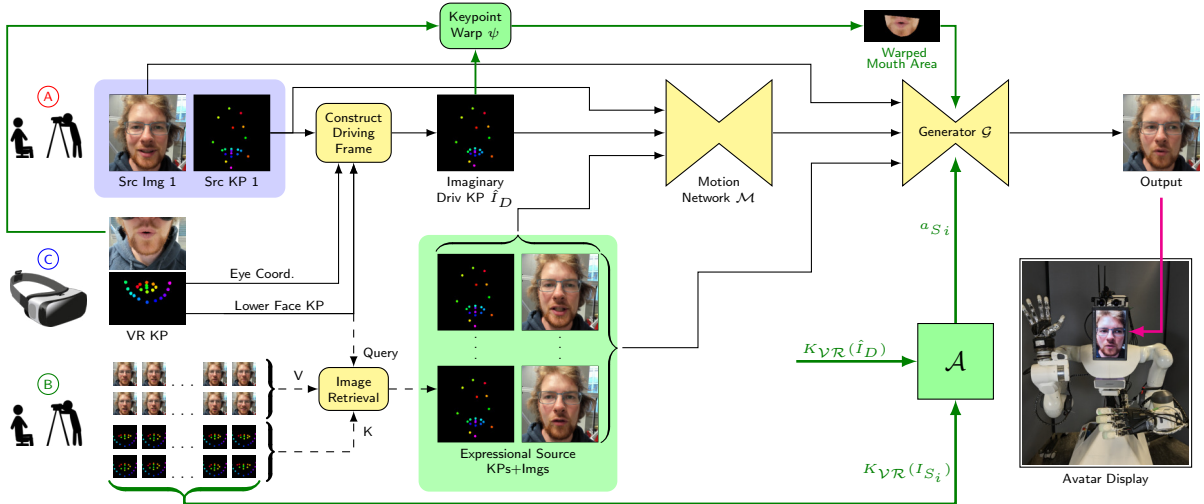
Making the operator seen properly is challenging, since they are wearing a VR headset. Simply capturing and displaying a video stream would be possible, but falls behind video conferencing solutions as facial expressions are partially hidden and distorted. Instead, we reconstruct the operator’s face from video data captured by a mouth camera and two eye cameras mounted on the HMD (see Fig. 6). Unlike related methods [45, 46, 48], we do not require per-operator training for an unseen person. Instead, our method is trained on large speaking-head datasets and generalizes to new operators.

Figure 5 gives an overview of our pipeline. We first capture a source video of an operator without the HMD and another one from the mouth camera on the HMD. From the first source video, we select four different fixed source images and optimize a keypoint mapping which projects keypoints from the mouth camera into the first source image, taking into account facial deformations caused by the HMD. Given the projected keypoints and eye tracking results, we dynamically construct imaginary driving keypoints.

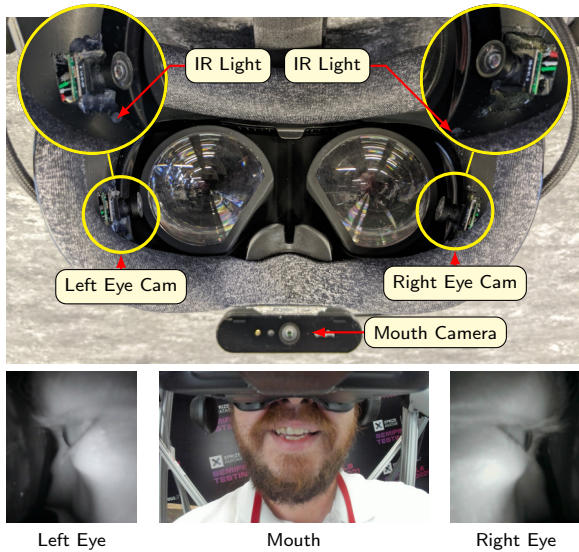
The encoded features of the source images and a dynamically retrieved expression image are deformed into the constructed driving keypoints using a deformation grid predicted by the motion network. We prevent temporal inconsistencies, which are mainly caused by expression image changes, with a source image attention mechanism and visual mouth camera guidance. Both are applied to the latent space before decoding the features to produce the output image. We will give an overview of these two innovations below. For a detailed explanation we refer to Rochow *et al.* [7].

#### Source Image Attention Mechanism

We select several different source images and train a source image attention mechanism that equips the network with the ability to decide how much information it requires from each source image. The attention values are then used to aggregate the latent representations of the source images after deforming the features to align them in the constructed driving keypoints. This significantly improves temporal consistency compared to our semifinal solution [52], as the attention values are estimated by a continuous function that smoothly



**Fig. 5** Inference pipeline for VR Facial Animation. New components compared to semifinals [52] are highlighted in green. We select 4-5 still source images from a portrait video of the operator shot before the run as source images (A). The remaining frames are optionally used as a key-value storage of retrievable expression keypoints and corresponding images (B). The live keypoints measured inside and outside the VR headset (C) are then projected to the first source image frame, where they are optionally used to retrieve the closest expression image with keypoints from the storage. The keypoints of all source images, including the retrieved one, and a constructed set of driving keypoints then enter the motion network  $\mathcal{M}$ , which estimates a warping grid that is used to deform the source images features, extracted by the generator-encoder network, to match the driving keypoints. The deformed features are aggregated over the source images in the lower facial area using a trainable attention mechanism  $\mathcal{A}$ . The mouth camera image from the HMD is warped into the lower facial area of the constructed driving keypoints and then encoded by a separate encoder network. An estimated mask gates the aggregated deformed source features using the warped mouth camera features. The masked aggregated features are then decoded to produce the output.



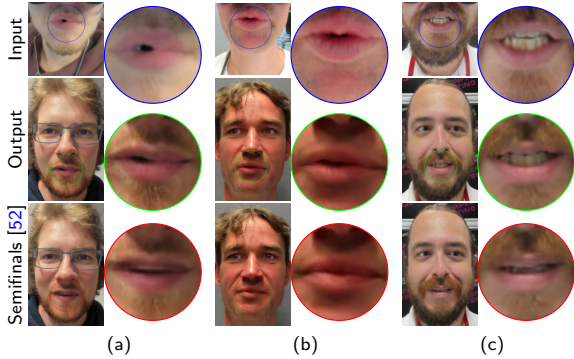
**Fig. 6** Modified Valve Index VR headset. We attached three additional cameras to capture the eyes and the mouth expression of the operator. We show the corresponding camera views at the bottom. Source: Rochow *et al.* [52].

adapts to changes in the mouth camera stream. Our previous approach, however, only utilized one retrieved expression image that can change abruptly and therefore introduce strong discontinuous effects. More diverse facial expressions of an operator are presented to the network which reduces the network’s dependence to the retrieved expression image and improves animation quality.

### Visual Mouth Camera Guidance

Including direct visual information from the mouth camera allows to resolve keypoint ambiguities and contains a broader range of facial expressions. However, it is difficult to train this part since large-scale datasets with mouth camera & full face images are not available (consider that the VR headset also visibly deforms the face, so a simple crop does not suffice).

We enable visual mouth camera guidance by estimating a Delaunay triangulation in the lower facial keypoints and using the barycentric coordinates to sample the mouth camera image in the lower facial area of the driving keypoints. This roughly aligns both representations. A trainable



**Fig. 7** Qualitative facial animation results. We show mouth camera input, the animated face, and results obtained by our semifinal solution [52] for three cases: (a) Transfer from a different operator in a particularly challenging case with sticking lips, (b) slightly open mouth, and (c) teeth showing.

encoder network then estimates a latent representation and conditions the aggregated deformed source image features via gated convolutions [54]. The masking operation in a gated convolution allows the elimination of poor image features that do not correspond to visual mouth camera information, while still encoding additional information in the latent representation. Another important advantage is that direct information propagation is prevented, which would lead to the network “pasting” the mouth section without correction. This allows us to continue training with entire faces. During training, we utilize the lower facial area of the driving image as the mouth camera input and simulate the imperfect perspective transformation by adding keypoint noise in both the source and target keypoints. In addition, different types of image noise are added to account for different lighting conditions. This, combined with the source image attention mechanism, already improves performance compared to our semifinal solution (see Semi vs. Ours-NF in Tab. 2). However, even when we emulate the effect of directly transforming the mouth camera keypoints to the driving keypoints, there are still differences that limit performance. We address this by manually searching for correspondences between mouth camera images and entire faces to annotate a small, suitable VR dataset consisting of 13 different persons. During finetuning, we select samples from this annotated set with a probability of 6%. This significantly improves performance.

**Table 2** Facial animation ablations

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	TIC $\downarrow$
Ours	<b>28.75</b>	<b>0.8586</b>	<b>0.0361</b>	+8.1%
Ours-5-Fix	28.50	0.8504	0.0376	+0.0%
Ours-NF	27.20	0.8369	0.0465	+3.7%
Semi [52]	25.10	0.7809	0.0580	+98.9%

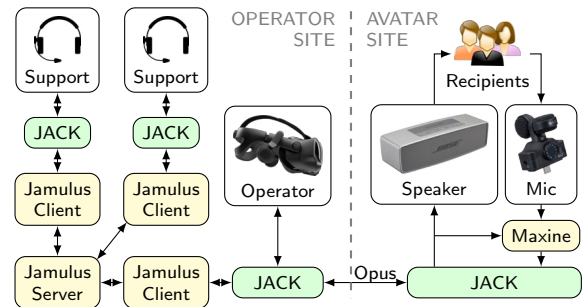
*Ours-5-Fix*: only five fixed source images without image retrieval. *Ours-NF*: No finetuning on manually annotated mouth camera images. *Semi*: Our semifinal solution. TIC: temporal inconsistency measure normalized to Ours-5-FIX. For details, we refer to [7].

### Facial Animation Evaluation

We evaluate our method variations and compare them to our previous method [52] on an annotated dataset of five unseen individuals, in which we manually assign HMD mouth camera images to roughly corresponding facial images. Accuracy and temporal inconsistency are reported in Tab. 2. The results indicate that all our method ablations outperform our previous work in terms of temporal consistency and accuracy. The highest accuracy is achieved when the last expression image is dynamically retrieved from the source video during inference (Ours). Temporal consistency, however, is maximized when all five source images are fixed during inference (Ours-5-Fix). This highlights the trade-off between accuracy and temporal consistency. We also show qualitative examples in Fig. 7.

### 4.4 Audio

Auditive perception and communication capabilities are key modalities for an immersive telepresence experience. A central objective for the design of the audio system is to optimize for low latency while maintaining the integrity of high-resolution stereo audio. Figure 8 shows an



**Fig. 8** Audio system. The setup allows multiple support crew members to listen in and to communicate with both the operator and recipient(s).

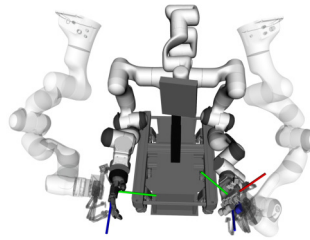


overview of our audio solution. The audio hardware on the operator station is comprised of the built-in microphone and headphones of the HMD, while the avatar robot is equipped with a stereo microphone mounted on top of its head display and a loudspeaker attached to its torso (see Fig. 1). The directionality of both devices on the avatar side, combined with the rigid connection between the microphone and the 6D movable head, establish a human-like experience w.r.t. room acoustics for both the operator and the recipients. All audio devices are connected utilizing the JACK audio connection kit<sup>3</sup>, matching the requirements for high resolution (24 bit & 48 kHz) and low latency (512 samples), while also providing the flexibility to establish robustness through layers of self-recovery, monitoring, and control.

The most expensive step in terms of latency is the WiFi transmission to and from the avatar robot. Therefore, we encode all audio packages using the OPUS audio codec and transmit them via UDP redundantly over the 2.4 GHz and 5.0 GHz WiFi networks (see Sec. 7). Depending on the concrete network latency and the set volume of the avatar’s loudspeaker, the operator can be exposed to hearing an echo of their voice feeding back from the loudspeaker to the microphone. To combat this, we integrate an echo cancellation system based on NVIDIA Maxine<sup>4</sup>, which also allows reduction of noise and reverberation. Finally, we integrate Jamulus<sup>5</sup> to provide audio conferencing functionality, augmenting the communication between the operator station and the avatar robot. In particular, each team member can wear headphones and join the audio conference to communicate with each other, with the operator, and with recipients through the avatar. This makes it easy for everyone involved to communicate and keep track of the current status, which is particularly helpful during setup, training, and monitoring.

## 5 Telemanipulation

Telemanipulation allows the operator to interact with remote environments and is a key element of our avatar system. The operator station and



**Fig. 9** Kinematic arm configuration for both the avatar robot (solid model) and the operator station (transparent model). The axes represent the common hand frame.

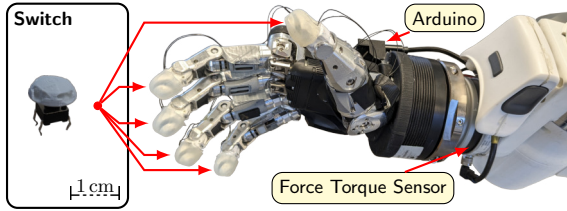
the avatar robot each use two Franka Emika 7 DoF Panda arms for this purpose. The operator wears SenseGlove DK1 hand exoskeletons which are mounted via Nordbo NRS-6050-D80 force-torque sensors to the Panda arms (see Fig. 1). This allows to track operator finger and hand positions and provides force and haptic feedback (see Sections 5.1 and 5.2).

The arms of the avatar robot are mounted in an anthropomorphic configuration to match the human arm workspace as closely as possible while minimizing the shoulder width to 78 cm, allowing for easy navigation through narrow passages (see Fig. 9). Since the Panda arms are neither symmetrical nor available in a mirrored version, the right hand is mounted in a 90° angle to avoid reaching the limit of joint five during normal manipulation tasks. OnRobot HEX-E force-torque sensors are mounted on the arms that hold a Schunk SVH and SIH hand on the right and left sides, respectively (see Figs. 10 and 14). Having two different hands on the avatar robot increases the manipulation capabilities for the operator: The active 9 DoF SVH hand allows very dexterous manipulation but has a rather low payload of about 1.5 kg. In contrast, the cable-driven 5 DoF Schunk SIH is advantageous for less precise but more forceful tasks. We modified all four Panda arms both in hard- and software to support our requirements: The firmware was customized to allow non-horizontal mounting. In addition, the modified firmware allows to automatically recover from error states under supervision of the control PC. This is one crucial feature that greatly increased our system robustness (see Sec. 8.2). In addition, we decreased the size of the avatar wrists by removing unused buttons, 3D printing smaller covers for the last joint, and mounting the teach buttons to a different location. This reduces

<sup>3</sup><https://jackaudio.org>

<sup>4</sup><https://developer.nvidia.com/maxine>

<sup>5</sup><https://jamulus.io>



**Fig. 10** Right Schunk SVH hand with custom fingertips holding pushbutton switches for contact measurements.

collisions, e.g. when manipulating objects on a table.

More details about the arm and hand controllers as well as our methods for providing force and haptic feedback are provided in the following.

## 5.1 Arm Control & Feedback

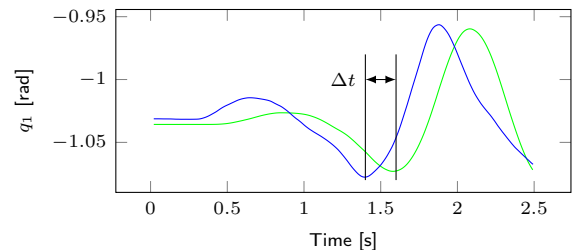
The avatar system utilizes two different arm controllers (operator station & avatar robot), which send joint torque commands to the corresponding Panda arms. Any information exchanged by both controllers (goal poses and force-torque measurements) are first transformed into a common frame located in the palm of the hands (see Fig. 9). This allows different kinematic chains for the avatar and operator station without any specific re-targeting. The control loop of the whole system runs at 1 kHz. At the operator station, the arm controller serves several purposes: First, it measures the 6D human hand pose and sends it to the avatar robot. Next, interaction forces measured on the avatar side are displayed to the operator. All hand movements are measured by the force-torque sensor mounted between the hand exoskeleton and the Panda arm. The controller uses these measurements to guide the arm following the human movement, generating a weightless feeling for the operator when no force feedback is displayed. Finally, the arm controller pushes the Panda arm away from any joint position or velocity limits. This is important to prevent the Panda arms from deactivating themselves, as humans can move their arm faster and have a larger workspace. Avoiding these limits on the operator side is straightforward. However, limiting the operator input to avoid joint limits on the avatar side, considering different kinematic chains, is not trivial due to latency constraints. Therefore, we implemented a model-based predictive limit avoidance module that predicts the avatar arm movements

based on the current joint state and the target pose commanded by the operator. This way, we can avoid joint limits from the avatar side by displaying forces to the operator. In addition, in case the operator overcomes the limit forces, the Panda arm will stop (see Sec. 8.1) and will be safely restarted (see Sec. 8.3).

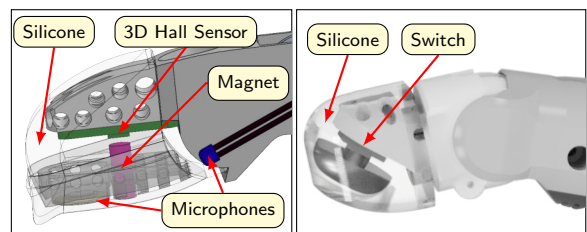
Besides our evaluation at the ANA Avatar XPRIZE Competition (see Sec. 9), we evaluated subcomponents of our telemanipulation arm controller. Fig. 11 shows the measured and predicted joint position of the first right arm joint during a teleoperated grasping motion. The prediction compensates the delay, which allows for instantaneous feedback of the avatar arm limits to the operator. We refer to Lenz & Behnke [4] for more details about the telemanipulation controller and component-wise evaluation.

## 5.2 Hand Control & Haptics

The hand controller maps the captured human finger positions to joint position commands for both Schunk hands. Different mappings are needed for the left and right hand (Schunk SIH and SVH). The SenseGlove DK1 measures a total of 20 DoF,



**Fig. 11** Predictive avatar model: Measured joint position for the first joint of the right avatar arm during a grasping motion (green) and predicted joint position for predictive limit avoidance (blue). Both measurements are captured on the operator side. Communication between both systems and motion execution generate a delay of up to 200 ms ( $\Delta t$ ), which is compensated by the predictive model.



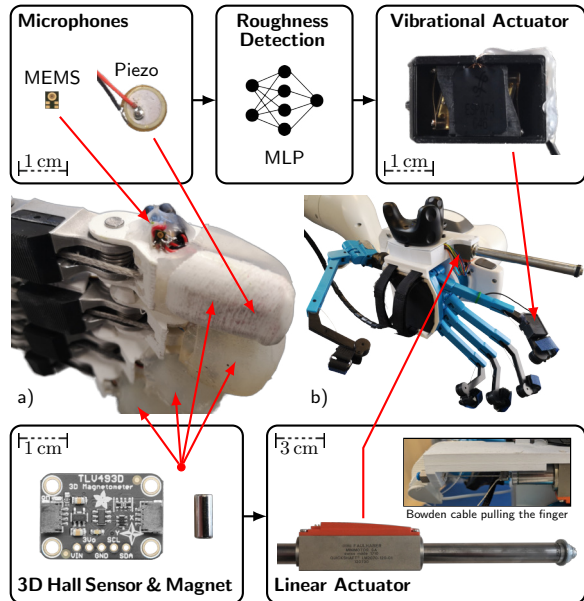
**Fig. 12** Fingertip sensors: SIH (left) and SVH (right).

4 joint angles per finger. In addition, a normalized flexion value is provided, indicating the total flexion of a particular finger. For the left SIH hand the finger flexion value is used to control the four finger flexion joints (ring finger and pinky are controlled by the same actuator). The thumb opposition is controlled directly by the corresponding joint measurement from the SenseGlove. The right Schunk SVH hand has nine actuators and therefore requires more fine-grained joint position commands. A joint-to-joint mapping is used for the SVH hand using the corresponding SenseGlove joints. The finger spread is controlled by calculating the angle between the index finger and pinky.

Providing per-finger haptic feedback to the operator when the avatar makes contact with objects or the environment is important for improved telemanipulation. Both Schunk hands provide motor currents and the measured joint angles. These measurements can be used to estimate contact or grip forces during active grasping actions. However, the contact between a finger and the environment when the finger is not actively moving is not visible in the data provided, because most of the finger joints are not backdrivable (SVH hand) or underactuated (SIH hand). To overcome the lack of information, we designed custom fingertips with additional sensors to replace the original ones.

Each fingertip (except the thumb due to space constraints) on the left SIH hand is equipped with an Adafruit TLV493D 3D magnet hall sensor and a small magnet embedded in a flexible silicon layer (see Figs. 12 and 13). Any contact acting on the fingertip moves the magnet and thus changes the magnetic field measured by the hall sensor. All hall sensor measurements are collected by a XIAO RP2040 microcontroller with a rate of 400 Hz and sent to the control PC. The SenseGlove vibration actuator is triggered for 200 ms if the absolute magnet field deformation exceeds a predefined threshold. This gives the operator a brief haptic feedback whenever the fingertip measures any contact.

The right Schunk SVH fingertips do not support the integration of similar Hall sensors due to their small size. Instead, we integrated 3D push-button switches providing binary feedback (see Figs. 10 and 12). Again, the SenseGlove vibration



**Fig. 13** Hardware implementation for roughness sensing and haptic feedback. a) Instrumented index finger on Schunk SIH hand. b) Instrumented index finger on SenseGlove DK1 hand exoskeleton.

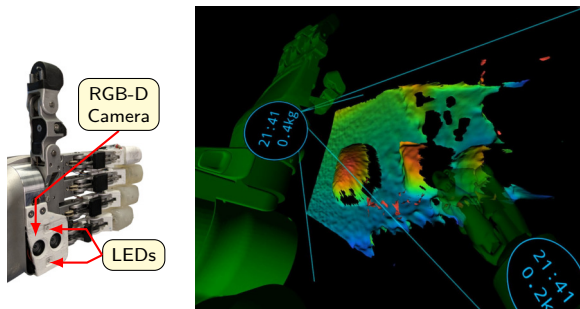
actuator is triggered for 200 ms when the contact switch is activated.

The SenseGlove contains active brakes that can prevent the operator from closing a particular finger. We activate the brake when the motor current from the corresponding avatar finger exceeds a predefined threshold. Both feedback modalities contribute to an immersive feeling when manipulating in the remote environment.

The ANA Avatar XPRIZE finals required the operator to distinguish different textures by touch alone. Besides a specific roughness sensing solution described below, we also replaced the left index finger brake with a Faulhaber LM2070 linear actuator (see Fig. 13). The actuator pulls the string connected to the SenseGlove fingertip link, enabling the system to actively extend the operator’s index finger. This allows the operator to feel contact forces even when the index finger is not actively moved.

### 5.3 Roughness Sensing & Display

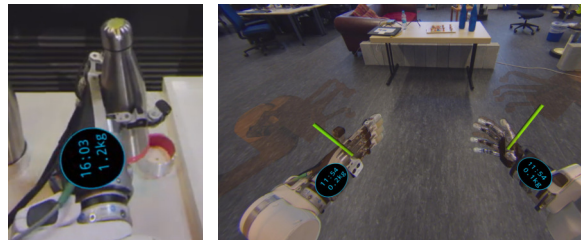
The index finger of the left Schunk SIH hand and its counterpart on the SenseGlove DK1 exoskeleton are equipped with a sensor and actuator setup (see Fig. 13), designed to let the operator



**Fig. 14** Solving the stone task without sight. Left: SIH hand equipped with RGB-D camera and LEDs. Right: VR visualization. Height is encoded as color (blue to red) and robot arms/hands are shown as a green overlay. The operator can fixate the stone using the right hand, freeze the view using a left thumb gesture, and then touch the stone with the left index finger to feel the texture.

intuitively discern between contacts with rough and smooth surfaces [6]. This low-cost and non-invasive approach uses two microphones capturing vibrations in the finger and the air around it, respectively. Very short sections ( $\sim 10$  ms) of these audio streams are then classified by a CNN as either rough or smooth. Finally, the classification results are used to modulate the frequency and amplitude of an oscillator, generating haptic signal that is directly fed to a vibrational actuator, capable of reproducing a spectrum of frequencies and amplitudes with fast response. We display rough surfaces by a 60 Hz sine wave with high amplitude and smooth surfaces by a 120 Hz sine wave with low amplitude. We note that the real-time nature of this system means that rough patches (bumps) are felt by the operator, while giving a slight buzzing sensation on other surfaces. As the approach is fully based on audio hardware, we integrate it with the rest of the audio system (see Sec. 4.4). This integration also allows us to inject the audio captured by the contact microphone inside the finger into the operator headset, invoking a more realistic and complete haptic perception by hearing the scratching sounds as well.

Due to the requirements specified in the competition rules (see Sec. 9), we designed the system to allow roughness sensing without direct sight. In this scenario, it is difficult for the operator to even find the objects in order to touch them. For this reason, we developed a 3D visualization based on geometry captured by a depth camera mounted in the left of the left SIH hand (see Fig. 14). It allows



**Fig. 15** Left: Wrist watch VR overlay shows current time and estimated weight. Right: If operator and avatar arm poses differ, a fade-in sequence is initiated. Rendered overlays show the operator arm pose, i.e. where the avatar arms will move to once the system is activated.

the operator to locate objects, hold them in place using the right hand, and move the instrumented index finger of the left hand over them. In the competition, this visualization was not required since the operator could see the stones and was asked to judge whether haptic feedback allowed to discern smooth and rough surfaces.

## 5.4 Status Visualization

Situational awareness for both the remote environment and the system status is important for successful teleoperation. Most of the developed system components focus on presenting the remote environment as immersive as possible while appealing to multiple human senses. Providing any additional information such as system health and different sensor measurements should not break the immersive teleoperation experience of the operator. Therefore, we implemented VR overlays which display additional information in natural ways. The current time and payload are rendered on each arm as a virtual wrist watch (see Fig. 15).

The telemanipulation subsystem is complex and does not always operate as expected by the operator. Important error notifications are displayed to the operator in various ways: A pressed E-Stop on the avatar side results in a red view for the operator. If the operator exceeds the head arm workspace by moving the head too far, the view will fade to black and an error message will be displayed. Similar, if one of the arms cannot follow the operator’s movement due to safety stops, network problems, or simply because the system is not activated, colored arm models are shown (see Fig. 15) to indicate that the system is not following. To minimize operator distraction during

normal operation, system status indicators are not visible when everything is running correctly.

## 5.5 Calibration

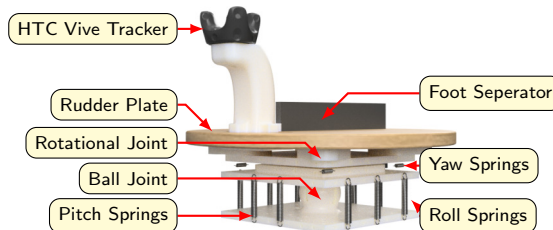
Before the avatar system can be used, multiple transforms and parameters need to be calibrated. We devised a principled approach starting at camera intrinsic calibration over hand-eye calibration on the robot side, to VR calibration on the operator side.

Intrinsic camera calibration is done using the kalibr software package [55]. The main vision cameras and ground cameras of the avatar robot have very high FoV with significant fish-eye distortion, thus the Double-Sphere camera model [56] is used to describe the intrinsics.

Extrinsic hand-eye calibration estimates the transformations between the cameras, the head arm, and the two main arms [3]. For this purpose, 3D-printed ArUco markers are mounted on the robot wrists. We use the ArUco marker detector of the OpenCV library to extract 2D pixel coordinates. During sample collection, the head continuously moves in a predefined sinusoidal pattern while the robot arm is moved manually using teach mode. Finally, the samples are used to compute optimal transforms using the Ceres solver.

The operator station is calibrated using the VR tracking setup. For this purpose, VR trackers are mounted on the exoskeleton wrists. After assembly, the arms are moved using teach mode and tracking poses are recorded. This allows precise estimation of the arm mounting poses and the operator station base pose relative to the VR coordinate system.

In addition to camera and robot calibration, the force-torque sensors at each wrist need to be calibrated [4]. Different end-effectors (SenseGloves, Schunk SIH, Schunk SVH hand, and corresponding 3D printed mounting adapters) result in different masses and center of mass. In addition, sensor bias results in barely usable raw sensor data. For calibration, 20 data samples from different sensor poses are collected. A standard least squares solver estimates the sensor parameters, i.e. the force and torque bias and the mass and center of mass of all attached components to compensate these effects. The calibration is performed once after every hardware change at the end-effectors or



**Fig. 16** Self-centering 3D rudder design with individually tunable springs for intuitive locomotion control.

if the bias drift is too large. This method does not compensate for bias drift online, but is sufficient for our application.

Finally, the nominal head pose is calibrated once the operator sits comfortably in the chair.

## 6 Locomotion

The omnidirectional base gives the operator advanced maneuverability. Four mecanum wheels with 8 inch (20.32 cm) diameter are driven by one RMD-X8 brushless motor each. The motor has a built-in 6.2:1 ratio gearbox and is connected via a 1:1 timing belt. To indicate the avatar’s movement to people in the remote environment, addressable RGB LEDs are mounted under the base plate. When the avatar is moving, the LEDs in the corresponding direction will light up. Running lights in clockwise or counter clockwise direction indicate rotation in place. In addition, the LEDs indicate a pressed E-Stop and the battery level during charging. A Raspberry Pi 4 Model B is used to communicate with all motors via a CAN interface, commands the avatar’s height (see Sec. 6.3), reads the current battery information, and sets the LED colors. The Raspberry Pi is connected to the main PC via Ethernet and runs a separate ROS instance. The base can be controlled using a standard wireless Xbox controller independent from the main PC.

### 6.1 3D Rudder

Providing locomotion control for a holonomic avatar robot platform in a VR setting is a challenging task. In our setup, control methods are constrained as the operator’s arms and hands control the bimanual arm and dexterous hands, and the VR headset pose controls the robot’s 6D-movable head. We propose a 3D rudder foot



**Fig. 17** Locomotion visualizations. Left: Birds-Eye view and predictions of base pose in the future. Right: Side-view for height adjustment.

input device with individually tunable springs for intuitive locomotion control (see Fig. 16).

The springs provide resistance and self-centering of the foot platform. The mechanical base of the rudder is built around a ball-bearing joint and a rotational thrust-bearing joint. Springs with different tension allow individual control of the resistance per axis. For absolute pose estimates, we attach an HTC Vive tracker to the rudder, which receives signals from the VR tracking system. The measured orientation (relative to the start orientation) is then translated to movement commands. We place a foot separator on the middle of the rudder’s surface to ease blind foot placement. In contrast to commercially available devices, our input device requires no calibration step by the operator.

Using the feet to pitch the rudder results in an intuitive control of the robot base for moving forward and backwards. Rolling the feet to the left and right allows for sideways control. Rotating the rudder on the yaw axis results in a rotation of the robot base. Mechanical end stops on the rotational axis prevent over-bending the springs.

## 6.2 Locomotion Visualizations

Despite the movable wide-angle cameras, the operator has limited view to the side and behind the robot. To overcome this limitation, especially for situation where the operator drives backwards, a rendered birds-eye view similar to a rearview mirror in a car is used to display additional information. The view slides down into the field of view when the avatar drives backwards or the operator looks upwards. Input for this comes from two Logitech Brio webcams with wide-angle converter that are mounted on the avatar’s upper body facing to the front and back of the robot. Fig. 1 shows

the location of the front camera. The rear camera is mounted in a similar way. The video streams are projected onto the ground plane using the camera calibration, stitched together with per-camera alpha masks, and displayed in the birds-eye view (see Fig. 17). The camera extrinsics were calibrated using a marker pattern laid in the area to the sides of the robot, which is visible in both cameras. The alpha masks are chosen in such a way that interference from e.g. the robot elbows is minimized. In some cases, parts of the elbow are still visible in the images, but this has not led to confusion of the operators so far. Furthermore, the projection and stitching assumes that all objects are on ground level. Violation of this assumption will lead to stitching errors. However, the boundary between the floor and any object will always be at the correct location. The “rear mirror” slides out of view when the operator stops driving and looks below the imaginary horizon for 1 second.

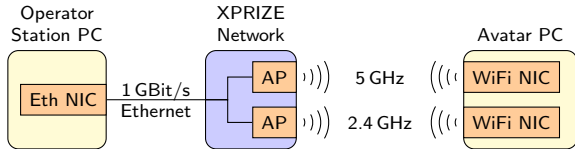
In addition, a 3D predictive model of the robot’s base is rendered while driving to facilitate anticipative navigation. Fig. 17 shows this base model in the VR view. The models display the base location in 2.5 and 5 seconds in the future, assuming constant velocity.

## 6.3 Avatar Height Adjustment

The upper body of our avatar robot can be adjusted in height to support manipulation at different heights and communication with standing and sitting persons. The linear axis adjusting the avatar’s height (shoulder heights reaching from 98 cm to 152 cm) is controlled using a bi-directional Danfoss KEP foot pedal (see Fig. 1). Tilting it forward will lift the robot, tilting it back will lower the robot. While adjusting the avatar’s height, the operator sees a rendered side view of the robot model, giving a better understanding of the current height (see Fig. 17).

## 7 Wireless Communication

It is clear that true avatar systems require freedom in mobility, unencumbered by cables. Our communication system makes use of two WiFi channels in the 2.4 GHz and 5 GHz bands, respectively (see Fig. 18). This allows to balance bandwidth across the bands and also to transmit some information



**Fig. 18** Network Architecture. The operator station contains a 1 GBit/s ethernet adapter, which is connected to the XPRIZE network (or our own AP during testing). Two separate access points broadcast a WiFi network at 2.4 GHz and 5 GHz, respectively. The avatar control PC is equipped with two WiFi adapters. Adapted from Schwarz *et al.* [5].

redundantly, increasing robustness to WiFi interference. We note that WiFi routers commonly offer dual-band operation.

The data streams and bandwidths are configured statically so that there are no bandwidth spikes at runtime which could lead to sudden WiFi saturation. The most bandwidth-heavy stream is caused by the main cameras on the robot head with  $2 \times 2472 \times 2178$  pixels @ 46 Hz. We use on-robot GPU-accelerated HEVC encoding and decoding to compress and decompress the data with minimal latency [5].

Tab. 3 shows the resulting data bandwidths and channel configurations. Manipulation control and audio data are transmitted redundantly over both channels, as they are particularly sensitive to packet drops.

## 8 System Safety, Monitoring & Robustness

Safety and system robustness is key when developing robotic avatar systems for both, the operator station and the avatar robot. Both systems are designed to directly interact with humans: The operator is strapped into the exoskeleton and the

avatar robot can touch and physically interact with recipients. In this section, we describe our safety measures, our monitoring tools that give the support crew situational awareness over the whole system at one glance, and system robustness procedures.

### 8.1 Safety Measures

Safety is the number one priority when building a robotic system that will be used by humans to interact with other humans. We use Franka Emika Panda cobot arms that are designed to work close to people. Their joint torque measurements per motor allow the controller to stop the arm immediately if an unexpected behavior occurs.

Two different E-Stops are integrated for both, the operator station and the avatar robot. On the operator side, the software E-Stop stops the arm controller and puts the Panda arms in teach mode. All motors will hold their position and the support crew can manually move the arm by pressing the teach buttons. A second E-Stop cuts the power to both arms, causing mechanical brakes to hold each joint in place.

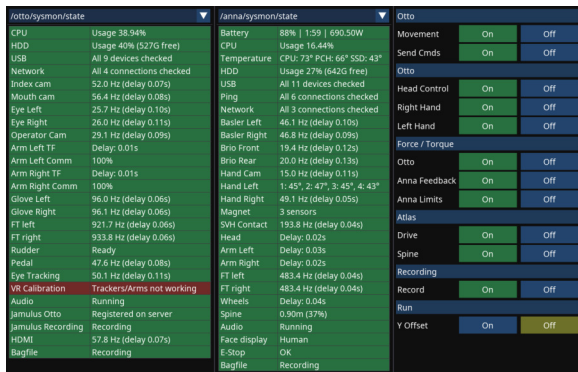
The same two E-Stops are integrated on the avatar robot. An HRI Wireless Emergency Stop serves as the software stop that puts the Panda arms in teach mode holding their current position. In addition, the head arm holds the current position and the avatar’s base is depowered, allowing a human to push the robot around. The hardware E-Stop is mounted on the avatar itself and cuts the battery power, resulting in a shutdown of the whole system including all motors, sensors, and the control PC.

### 8.2 System Monitoring

Monitoring is an essential part of robust robotics. It allows engineers to analyze problems and find their causes quickly. In our scenario, it was especially important to make sure the system is healthy before starting a run, since from then on, manual intervention was not permitted. During the run, the role of monitoring switches to a safety perspective, allowing the support crew to abort the run in case of danger to the human operator, the robot, or the environment. To be able to monitor the highly complex avatar system with one glance, we developed an integrated

**Table 3** Bandwidth requirements.

Downlink from avatar				Uplink to avatar			
Channel	MBit/s	5 GHz	2.4 GHz	Channel	MBit/s	5 GHz	2.4 GHz
Arm feedback	8.5	✓	×	Arm control	4.9	✓	✓
Transforms	4.1	✓	×	Transforms	1.4	✓	×
Main cameras	14.7	✓	×	Operator face	5.7	×	✓
Hand camera	5.5	×	✓	Audio	0.4	✓	✓
Diagnostics	0.4	✓	✓				
Audio	0.4	✓	✓				
Total [MBit/s]		28.1	6.3	Total [MBit/s]		6.7	11.0

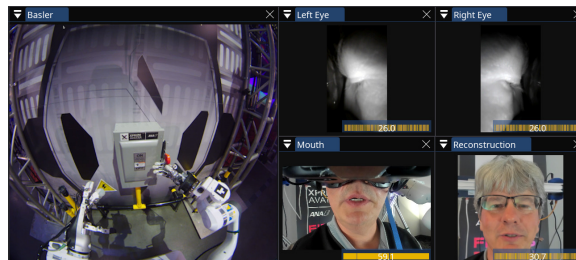


**Fig. 19** System Monitoring GUI. Left: Operator Station status. Each line corresponds to a system check. The red check indicates a problem with the VR trackers mounted on the exoskeleton—caused by a support crew member occluding the line-of-sight. Center: Avatar robot status. Right: Control buttons that enable/disable individual system components.

GUI. Because it contains a multitude of video streams and complex plots, the standard ROS GUI, `rqt`, was not suitable, as it is not optimized for high-bandwidth display. Instead, we developed a GUI based on `imgui`<sup>6</sup>, an immediate-mode GUI toolkit with OpenGL bindings. This allows us to decode and display the video streams directly on the GPU. The GUI follows the `rqt` paradigm with individual widgets that can be arranged via drag & drop.

The most important monitoring display is shown in Fig. 19. Both operator station and avatar robot run a `sysmon` node, which performs several checks with 1 Hz. These checks range from “Is hardware device X connected?” over “Does component Y produce data?” to “Is the operator station properly calibrated?”. The intention is simple: If all checks are successful, the support crew can start the run with confidence. Indeed, our policy was that every time an undetected error or misconfiguration led to a sub-optimal test run, a specific check for this condition was added. Overall, checks are similar to unit tests in software engineering, but monitor the live system in hardware *and* software.

Additionally, a section of the GUI with camera streams (see Fig. 20) together with headsets providing audio feedback give situational awareness to the support crew.



**Fig. 20** Camera streams. Left: Raw wide-angle camera stream (left eye) from the robot. Right: Eye cameras, mouth camera, and reconstructed animated face of the operator.

### 8.3 System Robustness

Ensuring support crew situational awareness and the connectionless network system are features that make the system more robust. However, there are many problems that can occur during a run, where manual intervention is not possible without aborting the trial. For this reason, we added auto-recovery mechanisms on multiple layers.

First, the Franka Emika Panda arms have independent safety systems which detect unsafe situations and either perform a soft-stop (braking with motor power) or hard-stop (engaging hardware brakes and switching off motor power). Since the operator can trigger both, e.g. by hitting an object with high speed, it is desirable to recover from these conditions. To this end, we modified the Panda firmware to be able to trigger recovery from an autonomous observer, which restarts the arms as long as the manual E-Stop is not triggered. During the restart of the arm, the operator is shown a 3D model of the arm to indicate that the arm is restarting and they should wait until the process is finished. The arm pose is then softly faded to the current operator pose and operation can continue [4].

Secondly, many hard- and software problems can be solved by simply restarting the affected processes [57]. As a simple example, restarting a device driver ROS node will recover after a transient disconnection of the device, without the need to make the driver node itself robust against such events. We stringently use the `respawn` feature of the ROS launch system to ensure that all nodes are automatically restarted whenever they exit. Furthermore, watchdog mechanisms are integrated that force nodes to exit which are stuck and do not produce output.

<sup>6</sup><https://github.com/ocornut/imgui>





**Fig. 21** Objects used in the ANA Avatar XPRIZE Semifinal scenarios: Solving a jigsaw puzzle (left), celebrating a business deal (middle), and exploring an artifact (right).

Finally, as a last line of defense, the main control PC is equipped with an external watchdog device. Our software running on the control PC regularly resets this watchdog. Should the system hang completely (which happened once during testing), the watchdog device will force a reset of the computer. Consequently, the software is configured to auto-start again, automatically resuming operations. The complete boot-and-recovery process takes less than one minute.

## 9 ANA Avatar XPRIZE Competition

The \$10M ANA Avatar XPRIZE competition challenged the robotics community to advance the state of the art in intuitive immersive telepresence systems [58]. The goal was to develop a robotic system that can transport human presence to a remote location in real time. A total of 99 teams registered in 2019 from 19 different countries across the world. One focus was pointed to intuitive and easy control of the avatar system. Thus, a panel of international experts with experience in related research fields were selected to evaluate all proposed systems at the semifinals and finals. In both events, one judge (the operator) controlled the avatar robot in a remote location solving manipulation and locomotion tasks, as well as communicating with a second judge (the recipient) through the avatar system. In this section, we present quantitative and qualitative results of our very successful participation in the semifinals and finals.

### 9.1 Semifinals

The semifinals were held in Miami, USA in September 2021 with 29 qualified teams. An additional 6 teams, unable to travel due to pandemic

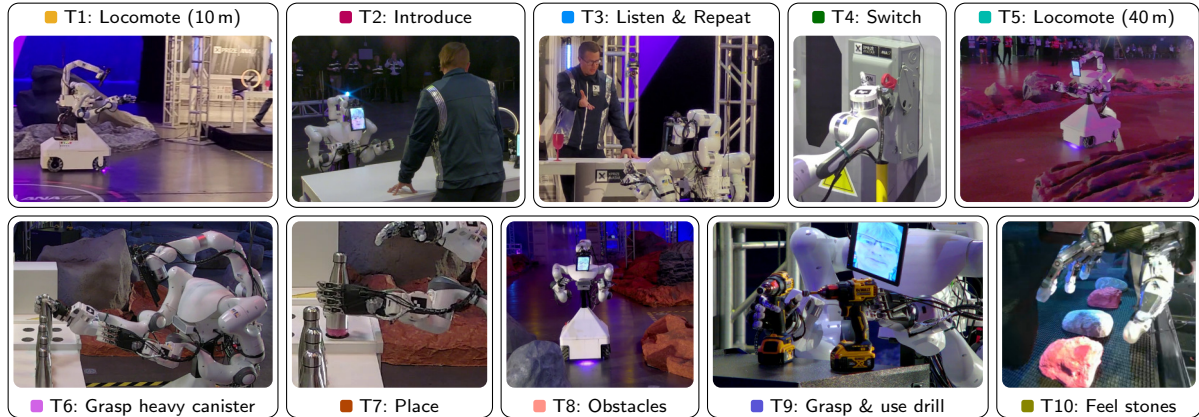
restrictions, were evaluated in their own labs in early 2022. All systems were evaluated through three different scenarios: collaboratively solving a puzzle, celebrating a business deal, and exploring an artifact. The scenario objects are shown in Fig. 21. All scenarios were tested on two days with different operator and recipient judges each. The best score per scenario over both days was included in the final score. A maximum of 30 points were awarded per scenario, based on operator experience (12 points), recipient experience (8 points), avatar ability (6 points), and overall system (4 points). In addition, 10 points were awarded based on a video where teams demonstrated their avatar system with self-chosen tasks in their own lab, resulting in a maximum score of 100 points.

The operator judge was located in the operator control room, separate from the scenario room where the avatar and recipient judge were located during the test runs. All communication between the two rooms had to go through the avatar system. Both the operator station and the avatar robot were allowed to be connected to wired network and power outlets. Teams had 60 min to train the operator. The operator then had up to 60 min to solve all scenarios. Tab. 4 shows the semifinal results for the top 20 teams qualified for the finals. Our system was ranked first with 99/100 points, only missing one point from the recipient experience of Scenario 1. We refer to Lenz & Behnke [4] for more in-depth analysis of our semifinal results.

### 9.2 Finals

The ANA Avatar XPRIZE finals took place in November 2022 in Long Beach, USA. A total of 17 qualified teams participated at the non-public qualification day. The top 16 and 12 teams advanced to the public testing on Day 1 and Day 2, respectively.

Similar to the semifinals, the operator controlling the avatar robot was located in the operator control room, which was separate from the arena where the test course was installed. This time, teams had to set up their operator station before each test in 30 min. Teams had 45 min in the operator control room to train and familiarize the operator with their system. In contrast to the semifinals, only one test course was available and therefore the operator training with the avatar



**Fig. 22** Tasks of the ANA Avatar XPRIZE Competition finals. ■ T1: Short locomotion (approx. 10 m) to the mission control desk. ■ T2: The operator introduces themselves to the mission commander. ■ T3: The operator receives mission details and confirms the tasks. ■ T4: Activate the power switch. ■ T5: Approx. 40 m of locomotion. ■ T6: Select a canister by weight (approx. 1.2 kg). ■ T7: Place the canister in the designated slot. ■ T8: Navigate around obstacles. ■ T9: Grasp and use the power drill to unscrew the hex bolt. ■ T10: Select the rough textured stone based on touch and retrieve it.

**Table 4** Results: ANA Avatar XPRIZE semifinals.

Rank	Team <sup>1</sup>	Scenario			Total <sup>2</sup>
		1	2	3	
1	NimbRo	29.00	30.00	30.00	99.00
2	iCub [59]	29.00	28.75	27.50	95.25
3	i-Botics [19]	28.25	28.00	27.75	94.00
4	Northeastern [17]	27.50	27.75	27.75	93.00
5	Dragon Tree Labs	26.75	28.50	27.75	93.00
6	AVATRINA [18]	26.50	28.00	28.25	92.75
7	Avatar-Hubo [22]	27.25	28.50	26.25	92.00
8	Converge	28.25	27.00	26.75	92.00
9	AlterEgo [21]	27.50	26.75	27.50	91.75
10	Cyberselves	26.25	28.00	26.50	90.75
11	Team SNU [20]	27.00	27.00	25.50	89.50
12	Pollen Robotics	25.75	27.75	26.00	89.50
13	Last Mile [23]	24.75	27.00	26.75	88.50
14	Enzo	25.00	27.00	25.25	87.25
15	Team UNIST	24.25	26.00	25.75	86.00
16	Inbioidroid	23.00	26.75	24.75	84.50
17	Rezilient	24.75	24.75	24.50	84.00
18	Touchlab	25.25	24.50	22.75	82.50
19	AvaDynamics	24.75	25.50	20.25	80.50
20	Janus	21.50	24.50	24.00	80.00

<sup>1</sup>Only the top 20 teams qualified for the finals are listed.

<sup>2</sup>Including 10 video submission points for each team.

took place inside the operator control room without the competition objects. Shortly before the competition run, the avatar robot was moved to the test course inside the arena where teams connected their system through the competition WiFi network provided by XPRIZE.

The operator judge had up to 25 min to complete the test course consisting of 10 tasks (see Fig. 22). The tasks had to be solved in order and included locomotion, communication with the recipient judge located in the arena, activating a

power switch, judging the weight of objects, using a power drill, and distinguishing a rough from a smooth textured stone. Teams were only allowed to interact with the operator judge after explicit requests by the judge.

### Final Results

The avatar systems were scored based on the task completion and the experience of the operator and recipient judges. Each successfully completed task was worth 1 point. The operator judge awarded up to 1 point each for the feeling of being present in the remote location, the ability to see and hear clearly, and the ease-of-use of the system. The recipient judge awarded up to 2 points for the first two criteria. Therefore, a maximum score of 15 points could be achieved. Ties were broken by completion time. The team’s final score was the better of the two competition days. Tab. 5 shows the final result for the 12 teams which advanced to Day 2 of the competition.

Our team NimbRo won the competition with a perfect score of 15 points. Our operator judges from both competition days were able to solve all ten tasks in 8:15 min and 5:50 min, respectively. In addition, our system received 5/5 judge points both days, resulting in two perfect runs with 15/15 points. Only Pollen Robotics were able to also receive a perfect score on Day 2 (they got 14,5 points on Day 1) but their operator needed almost twice as much time (10:50 min) to solve all ten tasks. Pollen Robotics’ and our system were

**Table 5** Results of the ANA Avatar XPRIZE Finals.

Rank	Team	Points			Time
		Task	Judge	Total	[mm:ss]
1	<b>NimbRo</b>	10	5.0	15.0	5:50
2	Pollen Robotics	10	5.0	15.0	10:50
3	Team Northeastern [17]	10	4.5	14.5	21:09
4	AVATRINA [18]	10	4.5	14.5	24:47
5	i-Botics [19]	9	5.0	14.0	25:00
6	Team UNIST	9	4.5	13.5	25:00
7	Inbiodroid	8	5.0	13.0	25:00
8	Team SNU [20]	8	4.5	12.5	25:00
9	AlterEgo [21]	8	4.5	12.5	25:00
10	Dragon Tree Labs	7	4.0	11.0	25:00
11	Avatar-Hubo [22]	6	3.5	9.5	25:00
12	Last Mile [23]	5	4.0	9.0	25:00

Only the top 12 teams advanced to Day 2 are listed.

the only ones solving all ten tasks on both competition days. Team Northeastern (placed 3rd) and AVATRINA (placed 4th) managed to solve all tasks on Day 2 and Day 1, respectively. All competing systems allowed the operator to complete the first four tasks on both days and also Task 5 on at least one day.

### Task Completion Times

We extracted the per-task completion times for both competition days from the official video feed<sup>7</sup>. Tab. 6 reports the task timings for all teams competing at the public competition days. Figure 23 compares the task timings for the six runs completing all ten tasks. Both of our competition runs were faster than any other successful run. As our operator judge on Day 1 solved all tasks in 8:15 min, giving us a comfortable lead over the other teams, our operator judge on Day 2 was instructed to take more risks by pushing our system to its limits. In addition, we greatly increased our avatar’s maximum base velocity for Day 2, resulting in much faster execution times for all tasks involving larger locomotion (Tasks 1, 4, 5, and 8) (see Tab. 6). We encountered a minor network issue during Task 9 on Day 1, which explains our longer execution time of 1:56 min, compared to 1:04 min on Day 2. All remaining tasks (2, 3, 6, 7, and 10) were solved within the same time ( $\pm 4$  seconds) on both days, showing the robustness of our system.

AVATRINA’s system had a much slower drive compared to the top three teams, as evidenced by slower execution times for the locomotion tasks.

The shorter tasks 1-3 (locomotion and communication with the recipient judge) and Task 7 (placing the canister into the designated slot) were consistently solved with similar execution times across the top six runs. Larger differences in individual task execution times are due to subsystem failures or sub-optimal grasp poses in case of the drill (Task 9). Pollen Robotics’ slower locomotion time (Task 5) on Day 2 was due to a reset of the operator control. AVATRINA had problems during the manipulation in Task 6, which resulted in a software restart on the operator side, costing 2:10 min. Both Pollen Robotics on Day 1 and Team Northeastern lost the first drill due to sub-optimal grasp poses. Both operators had to go back to the table and grasp the second drill before they could complete the task. Finally, Team Northeastern struggled to reach into the box on Task 10 while grasping the rough stone. Their avatar’s arm kinematics with the wrist above the hand resulted in collisions between the arms and the wall above the box. The left arm shut down completely during manipulation due to the collision and could not recover. However, the operator managed to retrieve the correct stone with the right arm after several attempts.

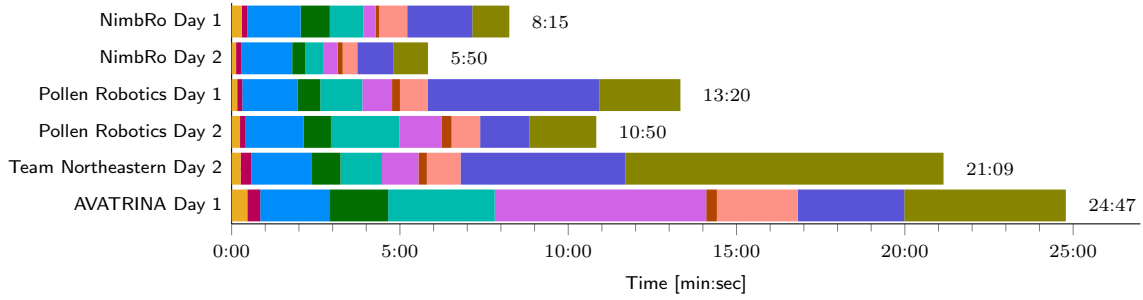
Some teams struggled starting their system in the arena environment using the competition WiFi—resulting in longer waiting times before the avatar began moving (see column “Start” in Fig. 23). For example, the merged team Cyberselves–Touchlab needed over 16 min to fix and reboot their system, which ended up missing to solve more than four tasks. This underlines the importance of robustness and ease of operation for such complex systems.

## 10 Operator Training

In the ANA Avatar XPRIZE Competition, an independent judge acted as the operator, who had never used the system before. Teams had only 45 min to train the operator and give them any advice necessary to successfully control the avatar. Thus, the training had to be optimized to provide enough information without overwhelming the operator. We present our operator training concept below.

It was crucial to have a clear plan in advance with a defined distribution of work among the team members. We made this plan about eight

<sup>7</sup><https://www.youtube.com/watch?v=lOnV1Go6Op0>



**Fig. 23** Per-task execution time for the top six competition runs solving all ten tasks. Tasks are color-coded as in Fig. 22.

**Table 6** Task completion times at ANA Avatar XPRIZE finals.

Team	Day	Time <sup>1</sup> [mm:ss]										Total	
		Start <sup>2</sup>	T1	T2	T3	T4	T5	T6	T7	T8	T9		T10
NimbRo	1	00:00	00:18	00:10	01:35	00:52	01:00	<b>00:22</b>	<b>00:06</b>	00:50	01:56	01:06	08:15
	2	00:00	<b>00:08</b>	<b>00:09</b>	01:31	<b>00:23</b>	<b>00:32</b>	00:26	00:09	<b>00:26</b>	<b>01:04</b>	<b>01:02</b>	<b>05:50</b>
1→2			0:00	-0:10	-0:01	-0:04	-0:29	+0:04	+0:03	-0:24	-0:52	-0:04	-2:25
Pollen Robotics	1	00:00	00:10	<b>00:09</b>	01:39	00:40	01:15	00:53	00:14	00:50	05:06	02:24	13:20
	2	00:00	00:15	<b>00:09</b>	01:43	00:49	02:02	01:15	00:18	00:51	01:28	01:59	10:50
Team Northeastern [17]	1	00:00	00:33	00:24	02:08	01:43	04:03	01:27	00:36	01:56			12:50
	2	00:00	00:16	00:19	01:47	00:52	01:14	01:05	00:15	01:00	04:54	09:27	21:09
AVATRINA [18]	1	00:00	00:28	00:23	02:03	01:45	03:10	06:17	00:19	02:24	03:10	04:48	24:47
	2	00:00	00:24	00:12	01:39	01:05	02:50	00:48	00:11	01:30	02:43		11:22
i-Botics [19]	1	00:00	00:13	00:26	<b>01:23</b>	01:53	01:57	01:52	02:07	02:57	09:47		22:35
	2	00:00	00:19	00:12	01:36	03:25							05:32
Team UNIST	1	00:00	00:25	00:33	01:55	01:15	02:38	02:51	00:25	01:29	08:06		19:37
	2	06:30	00:24	00:24	01:28	01:21	02:00	02:06	00:32	02:16	04:27		21:28
Inbiodroid	1	04:46	00:23	01:44	02:05	02:28	02:26	02:03	00:44	01:26			18:05
	2	05:10	00:40	00:35	02:04	03:09	02:40	04:10	00:27				18:55
Team SNU [20]	1	00:00	00:23	00:40	02:33	00:50	02:32	00:56	00:22	01:49			10:05
	2	00:00	00:26	00:18	01:36	01:26	02:39	00:32	00:30	01:49			09:16
AlterEgo [21]	1	00:00	00:50	01:08	02:22	04:18	03:18	10:21					22:17
	2	00:00	00:28	00:22	01:39	01:15	03:21	01:00	01:04	01:13			10:22
Dragon Tree Labs	1	05:27	00:13	00:17	01:33	05:30	01:50						14:50
	2	01:01	00:41	00:24	02:07	04:12	04:45	06:30	03:58				23:38
Avatar-Hubo [22]	1	00:00	00:23	00:30	01:44	00:37	07:11	03:17					13:42
	2	00:00	00:26	00:13	01:54	02:23	02:29						07:25
Last Mile [23]	1	00:00	03:00	02:42	01:27	03:43	04:13						15:05
	2	00:00	01:05	00:35	02:08	03:02	02:42						09:32
AvaDynamics <sup>3</sup>	1	02:50	03:19	01:06	02:11	11:39							21:05
iCub [59] <sup>3</sup>	1	00:53	01:09	01:12	02:08	04:28							09:50
Tangible <sup>3</sup>	1	02:25	00:15	00:09	01:57								04:46
Cyberselves – Touchlab <sup>3</sup>	1	16:10	00:26	00:11	01:33	04:29							22:49

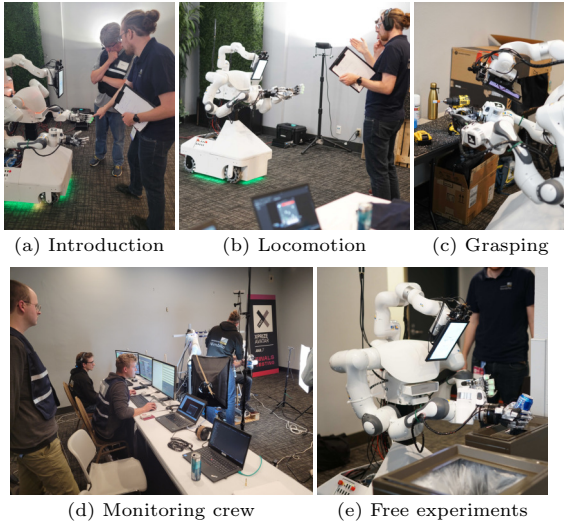
<sup>1</sup> The times were measured from the coverage of the competition: <https://www.youtube.com/watch?v=lOnV1Go6Op0>

<sup>2</sup> Time until the avatar started moving.

<sup>3</sup> Team did not advance to Day 2.

weeks before the final event and practiced the training procedure with external, untrained operators in our lab to get enough feedback and resolve missing details. Figure 24 shows images from our operator training from the second competition day.

In our team, one dedicated person was responsible for communicating with the judge and managing the training. This person was the judge’s direct contact for any questions. Next, we had one designated team member who monitored all software components and activated individual modules during the training. All software components



**Fig. 24** Operator Training. a) Introduction to the avatar robot. b) The operator receives instructions through the system and learns the locomotion capabilities of the system. c) Training to grasp and use the power drill. d) Crew monitors the training and starts individual software components. e) Let the operator playfully enjoy the system by solving competition-unrelated tasks.

necessary to run the system were started by this person to avoid any miscommunication. Our monitoring tools (see Sec. 8.2) were very helpful in providing at-a-glance system status. These two team members stayed in the operator control room during the competition run and were equipped with headsets that allowed them to listen to the audio communication between the operator and the avatar (see Sec. 4.4). During training and at the specific request of the operator judge, they were able to communicate through this audio channel. Note that all communications were audible to both the operator and the recipients on the avatar side. Therefore, the operator support crew could communicate directly with the avatar crew through the system, including the operator judge in any communication.

Next, we had a team member supporting the training by providing all necessary objects and monitoring the hardware of the system (Is the battery charged? Is a cable loose and needs to be fixed?, etc.). The fourth team member’s job was to set up the operator station, including the monitor setup, any cables that needed to be connected, and to launch the team communication components. The fifth person was responsible for any actions needed in order to start the facial animation (see

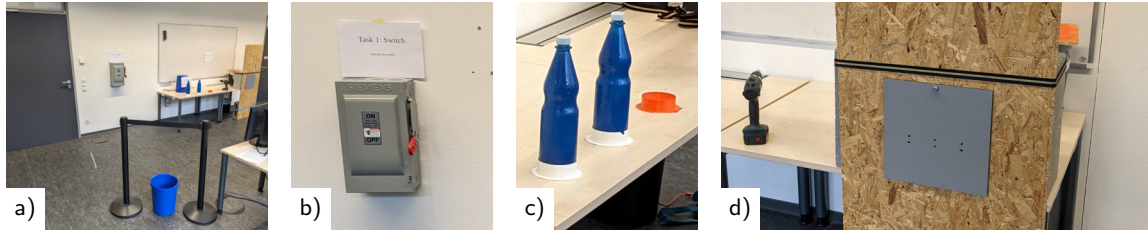
**Table 7** Operator training schedule.

Training	Time [min]
System overview	3
Face animation video w/o HMD	2
Put on HMD	1
Face animation video with HMD	2
Strap in hands	4
Enable arm and hand control	3
Locomotion training (T1, T5, T8)	4
Training switch and canister (T4, T6, T7)	5
Training power drill (T9)	5
Training stones (T10)	10
Enjoy the system	3
System recovery & recap	3
<b>Total training</b>	<b>45</b>

Sec. 4.3), including recording the necessary videos and feeding the data to the algorithm. The last two people acted as backups and were ready to solve short-term problems. Otherwise, they kept a low profile to not disrupt the process.

Our planned training schedule is summarized in Tab. 7. The training started with a brief overview of our system and a short safety briefing. We wanted the operator to feel safe using our system, which was achieved by providing information about implemented safety features (see Sec. 8.1). The overview eliminated some follow-up questions. Next, we recorded the videos (with and without HMD) needed for the facial animation method (see Sec. 4.3). The operator had to read a sentence which was placed next to the camera for the first video and displayed in VR for the second video. While the operator was strapped into the exoskeleton, the HMD displayed the image from the HMD camera facing the room in a look-through mode. This allowed the operator to see what was happening to their hands. Strapping the operator’s hands into the exoskeleton completed the control preparation, which took about 12 min.

For the next 30 min, the operator controlled the avatar and was trained to solve the competition tasks. First, individual subsystems were activated one at a time: Head movement, arm movement, finger movement, and finally locomotion control. After activating each subsystem, the operator explored their functionality briefly to get a good system overview (translational head movement to look around objects, birds-eye view for locomotion, force and haptic arm and finger feedback, etc.). Once the operator was able to control the avatar, the next step was to get used to



**Fig. 25** User study tasks. a) Test course overview including the barrier that had to be bypassed. b) Activating the switch (■ T4 in the competition). c) Selecting the heavy bottle and placing it into the orange ring (■ T6 & ■ T7). d) Using the power drill to unscrew the hex bolt and opening the hatch (■ T9).

the easier manipulation tasks (switch and canister). We used copies of the competition objects to give the operator the best possible training effect. Until now, the focus had been on the operator discovering the system. We supported the training by pointing out specific system features, mostly by asking the operator if they noticed them (i.e. Can you feel the force feedback? Can you inspect your hands from different angles?, etc.). For the more advanced tasks of using the power drill (T9) and feeling the stone texture (T10), we explicitly shared the strategies developed by our expert operators. Giving the operator the chance to explore different approaches might have led to a better understanding of the system’s capabilities, but was not possible due to the very limited training time.

An important aspect of operator training was to make it fun for the operator judge to control the system. To encourage this, the final training task was always a task not related to the competition. Examples include the operator throwing away a can or pressing the avatar’s own emergency stop. In addition, we encouraged the operator to look into a mirror and see their own animated face rendered on the avatar screen.

The training ended by unstrapping the operator from the exoskeleton before giving a short summary of the most important points and explaining the system recovery behaviors (see Sec. 8.2). The operator had a break of approximately 15 min while the avatar was moved into the arena. Right before the run, we strapped the operator back into the system, did final calibrations (head pose and eye tracking), and checked all system components.

All in all, our training preparation has been proven to be successful. Especially the testing of the training procedure in our own lab had solved important details in advance.

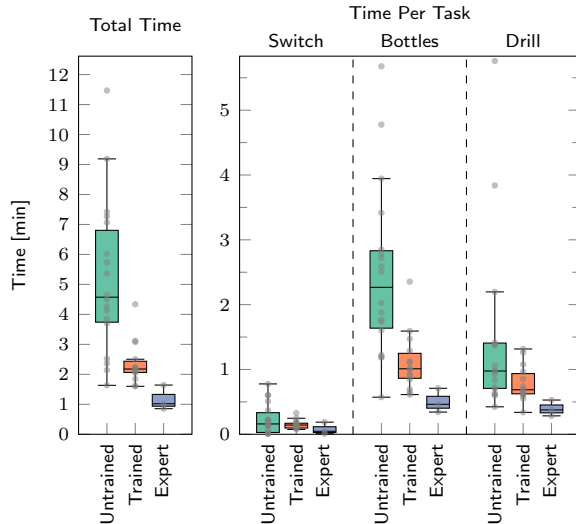
## 11 User Study

Developing an intuitive telemanipulation and immersive telepresence system for untrained operators was the main target of the ANA Avatar XPRIZE Competition. After the competition, we evaluated the intuitive control and usability of our system by conducting a user study. A total of 35 participants with an age of 20 to 34 years (average of 27.4 years) operated our avatar system and solved three tasks similar to the competition. Except for three of our team members, all remaining 32 participants had never controlled our system before.

We divided the participants into three groups: Untrained (18 participants), trained (14), and expert operators (three team members). All participants started by watching a short introduction video<sup>8</sup>, explaining how the avatar system works without giving any hints on solving the tasks. In addition, the “trained” participants were trained for 10 min to operate the avatar on the test course. Afterwards, all participants solved the three tasks on the test course using the avatar. The operator station and avatar robot were located in separate rooms approx. 30 m apart.

Fig. 25 shows the test course and the three tasks in detail. First, participants had to navigate around the barrier reaching and activating the switch, similar to Task 4 at the competition. Next, the heavier of two bottles had to be identified and placed inside the orange ring (Tasks 6 & 7 from the competition). Both bottles are painted and did not give a visual clue about their weight. Finally, similar to Task 9 from the competition, participants had to grasp and activate the power drill and use it to unscrew a hex bolt—opening a

<sup>8</sup>[https://www.ais.uni-bonn.de/videos/soro\\_2023\\_avatar\\_userstudy/](https://www.ais.uni-bonn.de/videos/soro_2023_avatar_userstudy/)



**Fig. 26** User study times. Left: Total mission execution time. Right: Times per task. Box plots show median and upper/lower quartile. Whiskers show extents disregarding outliers.

hatch. Some participants had difficulties grasping and holding the bottles. Similar to the competition, where multiple bottles were available, we gave participants multiple tries if they dropped the bottle out of reach of the avatar (i.e. laying under the table) by resetting the bottle to the original position. In case bottles fell over but were still reachable, they were not reset. The group of participants without training needed bottle resets in 7 out of 18 runs (39%). Participants with training dropped the bottles in two cases (14%). The expert operators did not drop any bottles.

All participants were able to solve all tasks within 12 min without any help. Figure 26 shows the time needed to solve all tasks and the time per task for all three study groups. As expected, trained operators were able to solve all tasks on average twice as fast as untrained operators (5:11 min and 2:22 min, respectively). Experts were again twice as fast as trained operators with an average time of 1:10 min to solve all tasks. The fact that the execution time for untrained operators only increases by a factor of two underlines the intuitive nature of the control, given the high system complexity.

Feedback from the participants after the study can be summarized as follows: The vast majority of the participants gave very positive feedback

about the immersive and intuitive control of the avatar. People often forget that their physical body is located in another room while they were controlling the avatar. Some people had problems with motion sickness. Most of them had similar experiences with other VR applications or reading in a car. The 3D ruder used to control the avatar’s locomotion could be improved with stronger springs and a visual overlay showing the current input values to provide a better awareness when people start driving. In addition, some participants had problems estimating the distance between the avatar and the table: These operators expected that their legs could collide with the table, since they were sitting. However, the avatar does not have legs. Clearer visual or auditory cues could resolve this uncertainty for future operators. Overall, people enjoyed operating our system and felt very present in the remote location.

## 12 Lessons Learned

Developing a complex robotic system and especially evaluating the system at a robotic competition gave insights and direct feedback on design choices that we would like to share.

### *Robustness*

Despite the extensive experience available in our group, the NimbRo avatar system is the most complex system we have ever built. Good monitoring, failure tolerance, and auto-recovery were extremely important for success. Our policy of adding a specific system check whenever a hardware or software problem occurred (see Sec. 8.2) saved a lot of time during development and especially during the competition, when every minute counted. In contrast to other teams, technical problems did not cause major delays or failures.

### *Frequent Testing under Competition Conditions*

Having a high test frequency allowed us to identify and address several points that were annoying or uncomfortable for operators, e.g. resulting in the improved rudder device. Frequent tests under competition conditions (untrained operator, time and space restrictions, and competition tasks) also helped the support crew to establish routine in efficiently training the operators as evidenced by the detailed routine described in Sec. 10.

### ***1:1 Correspondence is Best***

The connection between operator and avatar needs to be as close to identity as possible. Avoiding any scaling, offsetting, or 3D processing helps operators to quickly immerse into the system. In particular, correct hand-eye transformations let operators identify the avatar’s hands as their own.

### ***6 DoF Camera Motion***

The 6D motion of the wide-angle stereo camera mirroring the operator head movement greatly contributed to the immersion in the remote scene and allowed the operator to intuitively chose a viewpoint for manipulation that minimized occlusion. This especially became evident during the drill task, where operators could look from the side to see the trigger while grasping.

### ***Immersive Control Overlays***

One of our design criteria was to keep the VR overlays minimal and at places which are intuitive for untrained operators (i.e. the wrist watch and birds-eye view as a rear mirror). Complex overlays and VR menus distract the operator and limit the operator’s experience of being present in the remote space. Most of the operators who have controlled our system have reported immersive telepresence, which we attribute in part to the unobtrusive overlays mentioned above.

### ***Facial Animation and Gestures***

The photorealistic animation of the operator face on the avatar robot together with the realistic display movement and hand gestures enhanced the perception of the operator being present in the avatar robot.

### ***Operators differ***

Humans have different body proportions. Thus, the operator station needs to support a large variety of different operators (i.e. head size, finger, arm, and leg lengths, etc.). Our system needs only the initial head reference pose and the input to our face animation pipeline, allowing individual operator to use the system without much preparation or calibration. While this is true for the vast majority, we have seen problems with very short operators: Reaching the foot pedal and a smaller workspace due to shorter arms were the result.

Over 150 different operators tested our system as part of the user study (Sec. 11), during development, or during multiple system demonstrations. Very few operators complained about motion sickness. Most of them have similar problems with comparable VR applications. Individual operators have used the system for over 90 min at a time with no problems other than mild fatigue. We conclude, that our system is easy to use even during longer operation sessions, but removing motion sickness for every operator might be impossible.

### ***Modified Components***

Most of the components integrated in our avatar system are off the shelf components. We spent some effort in exploring the market and finding the best components. However, some important features were not available on the market. Therefore, modifying existing components to our needs gave us an advantage over other teams. Examples include the modified Panda firmware for non-horizontal mounting, additional eye-tracking and mouth cameras on the Valve Index HMD, and modified Schunk hand fingertips with additional contact sensors.

## **13 Conclusion**

This article presented the NimbRo avatar system, which won the \$10M ANA Avatar XPRIZE competition. We describe in detail the avatar robot with a humanoid upper body mounted on a mobile base, and the operator station, which consists of arm and hand exoskeletons, an HMD, and foot pedals. We provide subsystem evaluations for key components and refer to [2–7] for more detailed component analysis. The robustness of the system, achieved through comprehensive monitoring tools and multi-level system recovery, has been a major contributor to our success.

An important focus of the system design was to provide an intuitive and immersive operator interface for both trained and untrained operators. Achievement of these design goals is demonstrated in the extensive analysis of the semifinals and finals in the ANA Avatar XPRIZE competition. Our avatar system allowed the briefly trained judge to complete all ten tasks in 5:50 min, almost twice as fast as the second placed team. Key improvements over our semifinal system [2] such as a new base design, a linear actuator to



adjust the torso height, haptic perception, monitoring tools, failure tolerance, and robust wireless communication enabled this success.

Operator training within the limited time frame of 45 min was one important aspect of the competition. We described our training approach and team member roles during training in detail.

In addition to the competition analysis, we evaluated our system in a user study. Untrained operators were able to solve three locomotion and manipulation tasks with and without 10 min of training in a short time. Training on the complex system only reduced the execution time by a factor of two, compared to completely untrained operators. Experts with many hours of experience were only twice as fast as briefly trained operators. These results underline the intuitiveness and easy usability of the avatar system.

### **Data availability**

The datasets generated and analysed during this study are available from the corresponding author on reasonable request.

### **Compliance with Ethical Standards**

Conflict of Interest: The authors declare that they have no conflict of interest.

## **References**

1. Krotkov, E., Hackett, D., Jackel, L., Perschbacher, M., Pippine, J., Strauss, J., Pratt, G. & Orłowski, C. The DARPA robotics challenge finals: Results and perspectives. *The DARPA Robotics Challenge Finals: Humanoid Robots to the Rescue*, 1–26 (2018).
2. Schwarz, M., Lenz, C., Rochow, A., Schreiber, M. & Behnke, S. *NimbRo Avatar: Interactive immersive telepresence with force-feedback telemanipulation* in *International Conference on Intelligent Robots and Systems (IROS)* (2021), 5312–5319.
3. Schwarz, M. & Behnke, S. *Low-Latency Immersive 6D Televisualization with Spherical Rendering* in *International Conference on Humanoid Robots (Humanoids)* (2021).
4. Lenz, C. & Behnke, S. Bimanual telemanipulation with force and haptic feedback through an anthropomorphic avatar system. *Robotics and Autonomous Systems* **161**, 104338 (2023).
5. Schwarz, M., Lenz, C., Memmesheimer, R., Pätzold, B., Rochow, A., Schreiber, M. & Behnke, S. *Robust Immersive Telepresence and Mobile Telemanipulation: NimbRo wins ANA Avatar XPRIZE Finals* arXiv preprint arXiv:2303.03297. 2023.
6. Pätzold, B., Rochow, A., Schreiber, M., Memmesheimer, R., Lenz, C., Schwarz, M. & Behnke, S. *Audio-based Roughness Sensing and Tactile Feedback for Haptic Perception in Telepresence* in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2023).
7. Rochow, A., Schwarz, M. & Behnke, S. *Attention-Based VR Facial Animation with Visual Mouth Camera Guidance for Immersive Telepresence Avatars* in *International Conference on Intelligent Robots and Systems (IROS)* (2023).
8. Oh, P., Sohn, K., Jang, G., Jun, Y. & Cho, B.-K. Technical Overview of Team DRC-Hubo@UNLV’s Approach to the 2015 DARPA Robotics Challenge Finals. *Journal of Field Robotics* **34** (2017).
9. Stentz, A., Herman, H., Kelly, A., Meyhofer, E., Haynes, G. C., Stager, D., Zajac, B., Bagnell, J. A., Brindza, J., Dellin, C., *et al.* CHIMP, the CMU highly intelligent mobile platform. *Journal of Field Robotics* **32**, 209–228 (2015).
10. Karumanchi, S., Edelberg, K., Baldwin, I., Nash, J., Reid, J., Bergh, C., Leichty, J., Carpenter, K., Shekels, M., Gildner, M., *et al.* Team RoboSimian: Semi-autonomous mobile manipulation at the 2015 DARPA Robotics Challenge finals. *Journal of Field Robotics* **34**, 305–332 (2017).
11. Schwarz, M., Rodehutsors, T., Droeschel, D., Beul, M., Schreiber, M., Araslanov, N., Ivanov, I., Lenz, C., Razlaw, J., Schüller, S., *et al.* NimbRo Rescue: Solving disaster-response tasks with the mobile manipulation robot Momaro. *Journal of Field Robotics* **34**, 400–425 (2017).
12. Rodehutsors, T., Schwarz, M. & Behnke, S. *Intuitive bimanual telemanipulation under communication restrictions by immersive 3D*

- visualization and motion tracking in *International Conference on Humanoid Robots (Humanoids)* (2015), 276–283.
13. Klamt, T., Schwarz, M., Lenz, C., Baccelliere, L., Buongiorno, D., Cichon, T., DiGuardo, A., Droeschel, D., Gabardi, M., Kamedula, M., Kashiri, N., Laurenzi, A., Leonardis, D., Muratore, L., Pavlichenko, D., Periyasamy, A. S., Rodriguez, D., Solazzi, M., Frisoli, A., Gustmann, M., Roßmann, J., Süß, U., Tsagarakis, N. G. & Behnke, S. Remote mobile manipulation with the Centauro robot: Full-body telepresence and autonomous operator assistance. *Journal of Field Robotics* **37**, 889–919 (2020).
  14. Darvish, K., Penco, L., Ramos, J., Cisneros, R., Pratt, J., Yoshida, E., Ivaldi, S. & Pucci, D. Teleoperation of Humanoid Robots: A Survey. *IEEE Transactions on Robotics* (2023).
  15. Schmaus, P., Leidner, D., Krüger, T., Schiele, A., Pleintinger, B., Bayer, R. & Lii, N. Y. Preliminary Insights From the METERON SUPVIS Justin Space-Robotics Experiment. *IEEE Robotics and Automation Letters* **3**, 3836–3843 (2018).
  16. Schwarz, M., Stückler, J. & Behnke, S. *Mobile teleoperation interfaces with adjustable autonomy for personal service robots in International conference on human-robot interaction* (2014), 288–289.
  17. Luo, R., Wang, C., Keil, C., Nguyen, D., Mayne, H., Alt, S., Schwarm, E., Mendoza, E., Padir, T. & Whitney, J. P. *Team North-eastern’s approach to ANA XPRIZE Avatar final testing: A holistic approach to telepresence and lessons learned* arXiv preprint arXiv:2303.04932. 2023.
  18. Marques, J. M., Patrick, N., Zhu, Y., Malhotra, N. & Hauser, K. *Commodity telepresence with the AvaTRINA Nursebot in the ANA Avatar XPRIZE semifinals in RSS Workshop Towards Robot Avatars: Perspectives on the ANA Avatar XPRIZE Competition* (2022).
  19. Van Erp, J. B., Sallaberry, C., Brekelmans, C., Dresscher, D., Ter Haar, F., Englebienne, G., Van Bruggen, J., De Greeff, J., Pereira, L. F. S., Toet, A., *et al.* *What comes after telepresence? Embodiment, social presence and transporting one’s functional and social self in Int. Conference on Systems, Man, and Cybernetics (SMC)* (2022).
  20. Park, B., Jung, J., Sim, J., Kim, S., Ahn, J., Lim, D., Kim, D., Kim, M., Park, S., Sung, E., *et al.* *Team SNU’s avatar system for teleoperation using humanoid robot: ANA Avatar XPRIZE competition in RSS Workshop Towards Robot Avatars: Perspectives on the ANA Avatar XPRIZE Competition* (2022).
  21. Lentini, G., Settimi, A., Caporale, D., Garabini, M., Grioli, G., Pallottino, L., Catalano, M. G. & Bicchi, A. Alter-Ego: a mobile robot with a functionally anthropomorphic upper body designed for physical interaction. *IEEE Robotics & Automation Magazine* **26**, 94–107 (2019).
  22. Vaz, J. C., Dave, A., Kassai, N., Kosanovic, N. & Oh, P. Y. *Immersive auditory-visual real-time avatar system of ANA Avatar XPRIZE finalist Avatar-Hubo in International Conference on Advanced Robotics and its Social Impacts (ARSO)* (2022), 1–6.
  23. Haruna, M., Ogino, M., Tagashira, S. & Morita, S. *Verification of augmented avatar prototype 2 for one-click operation. XPRIZE/AVATAR final, Japan’s highest ranking in INTERACTION (Information Processing Society of Japan)* (2023).
  24. Martins, H., Oakley, I. & Ventura, R. Design and evaluation of a head-mounted display for immersive 3D teleoperation of field robots. *Robotica* **33**, 2166 (2015).
  25. Zhu, D., Gedeon, T. & Taylor, K. *Head or gaze? Controlling remote camera for hands-busy tasks in teleoperation: a comparison in Conference of the Computer-Human Interaction Special Interest Group of Australia* (2010), 300–303.
  26. Agarwal, P., Al Moubayed, S., Alspach, A., Kim, J., Carter, E. J., Lehman, J. F. & Yamane, K. *Imitating human movement with teleoperated robotic head in International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2016), 630–637.
  27. Whitney, D., Rosen, E., Phillips, E., Konidaris, G. & Tellex, S. *Comparing robot grasping teleoperation across desktop and virtual reality with ROS reality in Robotics*

- Research* (eds Amato, N. M., Hager, G., Thomas, S. & Torres-Torriti, M.) (2020), 335–350.
28. Sun, D., Kiselev, A., Liao, Q., Stoyanov, T. & Loutfi, A. A new mixed-reality-based teleoperation system for telepresence and maneuverability enhancement. *IEEE Transactions on Human-Machine Systems* **50**, 55–67 (2020).
  29. Stotko, P., Krumpfen, S., Schwarz, M., Lenz, C., Behnke, S., Klein, R. & Weinmann, M. A VR system for immersive teleoperation and live exploration with a mobile robot in *International Conference on Intelligent Robots and Systems (IROS)* (2019).
  30. Hirche, S. & Buss, M. Human-oriented control for haptic teleoperation. *Proceedings of the IEEE* **100**, 623–647 (2012).
  31. Abi-Farraj, F., Henze, B., Werner, A., Panzirsch, M., Ott, C. & Roa, M. A. Humanoid teleoperation using task-relevant haptic feedback in *International Conference on Intelligent Robots and Systems (IROS)* (2018), 5010–5017.
  32. Bimbo, J., Pacchierotti, C., Aggravi, M., Tsagarakis, N. & Prattichizzo, D. Teleoperation in cluttered environments using wearable haptic feedback in *International Conference on Intelligent Robots and Systems (IROS)* (2017), 3401–3408.
  33. Balachandran, R., Ryu, J.-H., Jorda, M., Ott, C. & Albu-Schaeffer, A. Closing the Force Loop to Enhance Transparency in Time-delayed Teleoperation in *International Conference on Robotics and Automation (ICRA)* (2020), 10198–10204.
  34. Wang, H., Liu, P. X. & Liu, S. Adaptive neural synchronization control for bilateral teleoperation systems with time delay and backlash-like hysteresis. *IEEE Transactions on Cybernetics* **47**, 3018–3026 (2017).
  35. Panzirsch, M., Singh, H., Krüger, T., Ott, C. & Albu-Schäffer, A. Safe interactions and kinesthetic feedback in high performance earth-to-moon teleoperation in *IEEE Aerospace Conference* (2020).
  36. Guanyang, L., Xuda, G., Lingzhi, L. & Yan, W. Haptic based teleoperation with master-slave motion mapping and haptic rendering for space exploration. *Chinese Journal of Aeronautics* **32** (2019).
  37. Lichtenstein, L., Barabas, J., Woods, R. L. & Peli, E. A feedback-controlled interface for treadmill locomotion in virtual environments. *ACM Transactions on Applied Perception (TAP)* **4** (2007).
  38. Iwata, H., Yano, H., Fukushima, H. & Noma, H. Circulafloor [locomotion interface]. *IEEE Computer Graphics and Applications* **25**, 64–67 (2005).
  39. Bouguila, L., Evequoz, F., Courant, M. & Hirsbrunner, B. Walking-pad: a step-in-place locomotion interface for virtual environments in *Proceedings of the 6th international conference on Multimodal interfaces* (2004), 77–81.
  40. Ohshima, T., Ishihara, H. & Shibata, R. Virtual ISU: Locomotion interface for immersive VR gaming in seated position in *Proceedings of the 2016 Virtual Reality International Conference* (2016), 1–4.
  41. Carmichael, C., Balderas, M. V., Ko, B., Nova, A., Tabafunda, A. & Uribe-Quevedo, A. Spring Stepper: A Seated VR Locomotion Controller in *22nd Symposium on Virtual and Augmented Reality (SVR)* (2020), 346–350.
  42. Otaran, A. & Farkhatdinov, I. A short description of an ankle-actuated seated VR locomotion interface in *Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2021), 64–66.
  43. Kitson, A., Hashemian, A. M., Stepanova, E. R., Kruijff, E. & Riecke, B. E. Comparing leaning-based motion cueing interfaces for virtual reality locomotion in *IEEE Symposium on 3D user interfaces (3DUI)* (2017), 73–82.
  44. Chu, H., Ma, S., De la Torre, F., Fidler, S. & Sheikh, Y. Expressive telepresence via modular codec avatars in *European Conference on Computer Vision (ECCV)* (2020), 330–345.
  45. Lombardi, S., Saragih, J., Simon, T. & Sheikh, Y. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* **37**, 1–13 (2018).
  46. Wei, S.-E., Saragih, J., Simon, T., Harley, A. W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H. & Sheikh, Y. VR

- facial animation via multiview image translation. *Trans. on Graphics (ToG)* **38**, 1–16 (2019).
47. Olszewski, K., Lim, J. J., Saito, S. & Li, H. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (TOG)* **35**, 1–14 (2016).
  48. Richard, A., Lea, C., Ma, S., Gall, J., De la Torre, F. & Sheikh, Y. *Audio-and gaze-driven facial animation of codec avatars in Winter Conference on Applications of Computer Vision (WACV)* (2021), 41–50.
  49. Luo, R., Wang, C., Schwarm, E., Keil, C., Mendoza, E., Kaveti, P., Alt, S., Singh, H., Padir, T. & Whitney, J. P. *Towards Robot Avatars: Systems and Methods for Teleinteraction at Avatar XPRIZE Semi-Finals in Int. Conf. on Intelligent Robots and Systems (IROS)* (2022).
  50. Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R. & Ng, A. *ROS: an open-source robot operating system in International Conference on Robotics and Automation (ICRA) Workshop on Open Source Robotics* (2009).
  51. Dodgson, N. A. *Variation and extrema of human interpupillary distance in Stereoscopic displays and virtual reality systems XI* **5291** (2004), 36–46.
  52. Rochow, A., Schwarz, M., Schreiber, M. & Behnke, S. *VR Facial Animation for Immersive Telepresence Avatars in International Conference on Intelligent Robots and Systems (IROS)* (2022), 2167–2174.
  53. Triantafyllidis, E., Mcgreavy, C., Gu, J. & Li, Z. Study of multimodal interfaces and the improvements on teleoperation. *IEEE Access* **8**, 78213–78227 (2020).
  54. Dauphin, Y. N., Fan, A., Auli, M. & Grangier, D. *Language Modeling with Gated Convolutional Networks in International Conference on Machine Learning (ICML)* (2017).
  55. Rehder, J., Nikolic, J., Schneider, T., Hinzmann, T. & Siegwart, R. *Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes in IEEE International Conference on Robotics and Automation (ICRA)* (2016), 4304–4311.
  56. Usenko, V., Demmel, N. & Cremers, D. *The double sphere camera model in International Conference on 3D Vision (3DV)* (2018), 552–560.
  57. Candea, G., Kawamoto, S., Fujiki, Y., Friedman, G. & Fox, A. *Microreboot - a technique for cheap recovery in Symposium on Operating Systems Design and Implementation (OSDI)* (2004).
  58. Behnke, S., Adams, J. A. & Locke, D. The \$10 Million ANA Avatar XPRIZE Competition Advanced Immersive Telepresence Systems. *arXiv preprint arXiv:2308.07878* (2023).
  59. Dafarra, S., Darvish, K., Grieco, R., Milani, G., Pattacini, U., Rapetti, L., Romualdi, G., Salvi, M., Scalzo, A., Sorrentino, I., et al. iCub3 avatar system. *arXiv preprint arXiv:2203.06972* (2022).