Redirecting Desktop Interface Input to Animate Cross-Reality Avatars

Jason W. Woodworth David Broussard Christoph W. Borst

CACS VR Lab
University of Louisiana at Lafayette

Abstract

We present and evaluate methods to redirect desktop inputs such as eye gaze and mouse pointing to a VR-embedded avatar. We use these methods to build a novel interface that allows a desktop user to give presentations in remote VR meetings such as conferences or classrooms. Recent work on such VR meetings suggests a substantial number of users continue to use desktop interfaces due to ergonomic or technical factors. Our approach enables desktop and immersed users to better share virtual worlds, by allowing desktop-based users to have more engaging or present "cross-reality" avatars. The described redirection methods consider mouse pointing and drawing for a presentation, eye-tracked gaze towards audience members, hand tracking for gesturing, and associated avatar motions such as head and torso movement. A study compared different levels of desktop avatar control and headset-based control. Study results suggest that users consider the enhanced desktop avatar to be human-like and lively and draw more attention than a conventionally animated desktop avatar, implying that our interface and methods could be useful for future cross-reality remote learning tools.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction Paradigms—Virtual Reality; Human-centered computing—Interaction design—Interaction design process and methods—User interface design

1 Introduction

Online remote instruction during the COVID-19 pandemic highlighted the difficulty of remote meeting spaces to effectively replace in-classroom education. This contributed to many students making little progress while learning at home [6]. Some of this may be due to a lack of feeling present in a class or engaging with a teacher.

VR meeting spaces provide an immersive alternative for students, and VR may restore a sense of presence and belonging to a class [26]. Research on the role of immersed user avatars in educational VR has shown that having a teacher immersed with the students in a scene can help increase engagement and lesson progression [22]. However, VR interfaces impose challenges for teachers, including discomfort after repeated use. In recent real-world applications, users with the option to choose desktop or immersive VR meetings often switched to desktop after technical or ergonomic problems [1, 26].

VR meeting tools can provide "cross-reality" interfaces [14] that allow non-immersed users using standard desktop devices to interact with fully-immersed VR users, but desktop interfaces reduce avatar control. Whereas a standard tracked VR interface maps real-life head and hand motion directly to an avatar, desktop interfaces, such as those in Mozilla Hubs, VRChat, or Virbela, offer limited degrees of freedom and less direct control through keyboard and mouse interaction. This limits avatar expressiveness or involves canned idle animations that may not reflect the user's intent.

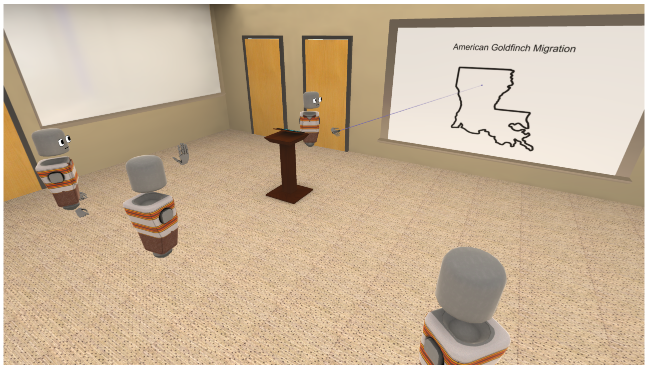

Figure 1: An example classroom environment. The teacher avatar points towards the board while students immersed with standard headset interfaces view the lecture.

The quality and expressiveness of avatars in collaborative VR has been found important for maintaining social presence and performance [24], and increased social presence has been linked to better performance in online education [8]. Prior work studying educational presentations in VR indicates that students find natural head and hand motions, ability to make eye contact, and an indication of talking to be important to their engagement [26].

We present a novel desktop interface that uses desktop-based head tracking, eye tracking, and hand tracking, in addition to mouse input, to redirect natural human motion onto a VR avatar (seen in Figure 1). Redirection methods are designed to allow the avatar to perform the most important teaching-related gestures more like a fully tracked VR avatar. This includes making eye contact with students and important objects, pointing at and drawing on a slideshow presentation, and making hand-based gestures such as waving or indicating size. To accommodate multiple redirection methods running at once, a priority-based system determines target poses for the avatar head, hands, and pupils to best represent the user's attempted expression. Finally, a body mechanics system moves the avatar's body parts to target poses with more realistic body motion.

We present an evaluation of these techniques, along with insight into the importance of an avatar's liveliness in an educational setting. The redirected avatar is compared to a fully tracked VR avatar, to a redirected avatar that does not take advantage of eye or hand tracking sensors, and to a more standard first-person controller avatar similar to what is used in software like Mozilla Hubs.

2 Related Works

Cross-Reality Interfaces "Cross-reality" is a term for interfaces that simultaneously straddle multiple points on the reality-virtuality continuum [12]. This can refer to asymmetric interfaces that enable interaction between users at different points on the continuum, so they do not have an equivalent experience across "realities" [21]. Recent research considers interaction techniques for collaborative cross-reality systems. For example, in [14], a pointing interface allows a non-immersed touchscreen user to point to a scene object, with the pointing presented to an immersed user in a manner tailored to the immersive environment. More generally, works such as [18] discuss guidelines for seamless integration between 2D and

This is an author-formatted version. Original publication: J. W. Woodworth, D. Broussard, and C. W. Borst, "Redirecting Desktop Interface Input to Animate Cross-Reality Avatars", IEEE Conference on Virtual Reality and 3D User Interfaces 2022; 12-16 March, Virtual

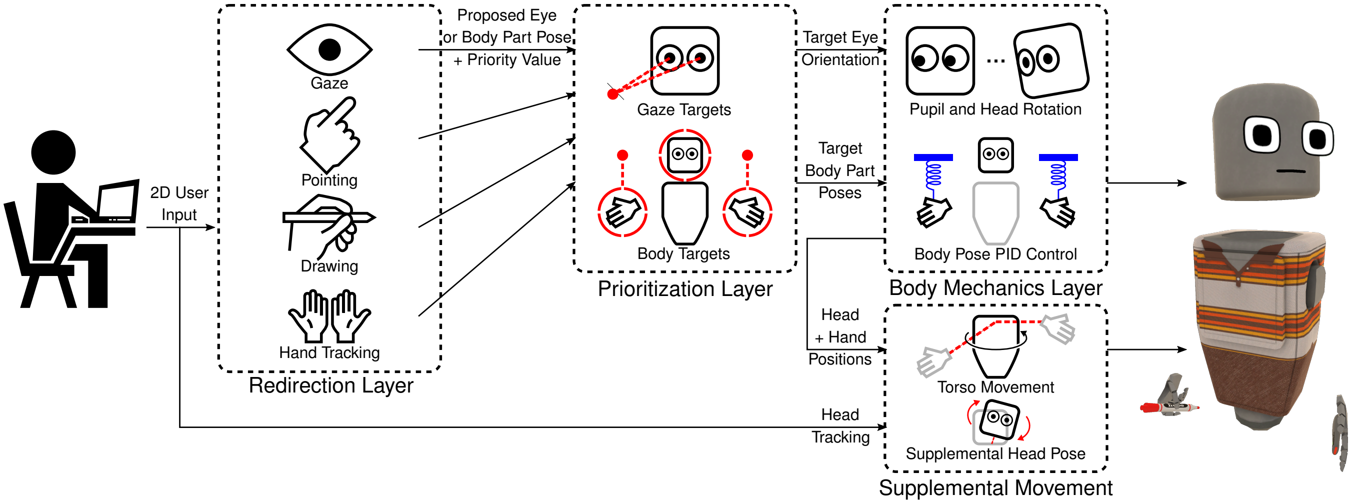

Figure 2: An overview of the components of the priority-based redirection system that animates the avatar.

VR interfaces, emphasizing the importance of consistent visibility of users and landmarks (such as screens) across interfaces.

Remote VR Meetings VR-based remote meetings were increasingly seen for conferences and meetings during the COVID-19 pandemic. However, recent applications of VR-based meeting spaces for education and conferences have raised concerns about the ergonomics of long-term headset use. For the IEEE VR 2020 conference [1], the VR platform Mozilla Hubs was ranked as providing the highest sense of social presence (compared to other desktop-based tools), but headset usage reduced over the course of the conference due to factors such as external keyboard use and overall VR fatigue. Another study conducted on a remote class [26], also using Mozilla Hubs, showed wide variation in headset ratings and mixed preference for headset-based and desktop-based class meetings.

Avatar and Gaze Redirection Redirection techniques have been used on 3D representations of 2D video streams, treating these like "avatars" and rotating them towards in-scene targets detected using desktop eye trackers [23]. Bailenson et al. discuss effects of a gaze manipulation that presented multiple immersed users with different gazes from a single avatar [3]. Subjects were unlikely to detect the manipulation. Gaze redirection has also been applied to asymmetric interfaces, as in [9], where high-level gaze target detection is input to a redirection that considers socially-motivated rules (e.g., determining dwell time limits and head motion speed).

Prior research on hybrid gaze (not fully natural or synthesized) [20] offers some insight into user preference for models based on social guidelines rather than randomized behavior. The work's authors note several limitations; the blending between real and simulated gaze was unsophisticated (and thus noticeable) and the experiment was not conducted in an immersive environment. Similar work on coordinating collaborative interaction [2] has explored how social cues are delivered and read through eye gaze, noting patterns of eye behavior in different phases of interaction.

Some methods of gaze presentation use additional abstract visual augmentation to reinforce phenomena such as joint attention [19], side-stepping the need to construct rules based on a traditional social understanding of gaze but forgoing the use of a humanoid avatar.

Some works redirect other motions such as gestures. The asymmetric AR system in [15] uses gaze and pointing redirection to redirect user motion to correspond to the locations of objects in the real world. The VR system in [17] animates gestures from real-time speech input based on recorded video of the presenter. Other researchers looked at improving gestural presentation, such as [16] evaluating the friendliness of gestures in an automated agent.

3 Redirection Techniques

3.1 Architecture Overview

Figure 2 overviews the architecture of our redirection system. User input is captured through a set of desktop devices and sent to individual redirection methods in the *Redirection Layer*. Currently, these devices include the mouse, a Tobii Eye Tracker 5 for eye and head tracking, and a Leap Motion Controller for hand tracking. Other desktop interaction devices, such as webcams for gauging facial expression and mouth movement, could be added.

Redirection methods take input from each device and, based on the specific method, produce target poses for avatar body parts. Target poses are fed to the *Prioritization Layer* along with a priority level value. Target poses are then sorted by priority, per body part, and the highest-priority poses are sent to the next layer. This allows certain redirection methods to interrupt others. For example, a gaze method may direct the avatar to look at a student with a priority of 50, then be interrupted by the user engaging a pointing method directing the avatar to look at the board with a priority of 75. A gaze method could then use priority level above 75 to move the gaze back to the student, for example, automating an occasional glance.

Target poses are sent to the *Body Mechanics Layer*, where each body part is animated to move towards its target pose. To avoid sudden discrete changes of simply changing to the target pose, body parts such as hands are driven towards targets by a PID controller, which gives more human-like continuous motion with acceleration and deceleration and possible small corrective motions. Head and pupil rotations are controlled by a model of human motion allowing the pupils to rotate without triggering head rotation unless certain thresholds are met (details are provided later).

A *Supplemental Movement* module sits on top of the Body Mechanics Layer and adds additional motion details to improve liveliness. For example, it can superimpose head motion tracked by the Tobii desktop Eye Tracker onto the final head pose. This allows the user to, for example, nod their head to respond to a student and have their natural motion be reflected by the avatar.

In our study of redirection, these methods animate a stylized robot presenter avatar modified from Mozilla Hubs. This avatar was chosen for its simplicity, to avoid problems with the uncanny valley, and modified with large eyes emphasizing gaze. Students in the environment animate the same avatar with a standard headset-based VR interface and are in front of the teacher (Figure 1).

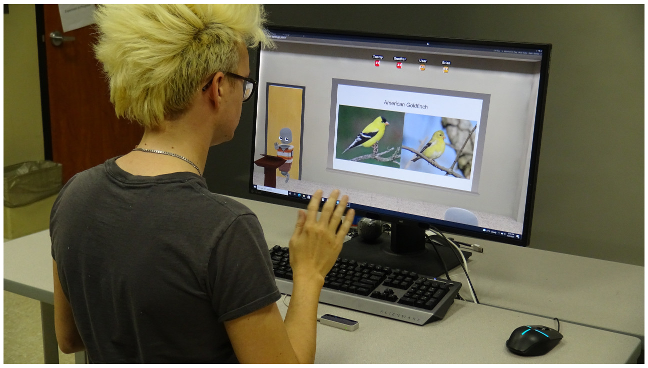

Figure 3: A user sitting at the teacher interface waving to students. A Leap Motion Controller is positioned on the desk directly in front of them. A Tobii Eye Tracker is attached to the bottom of the monitor.

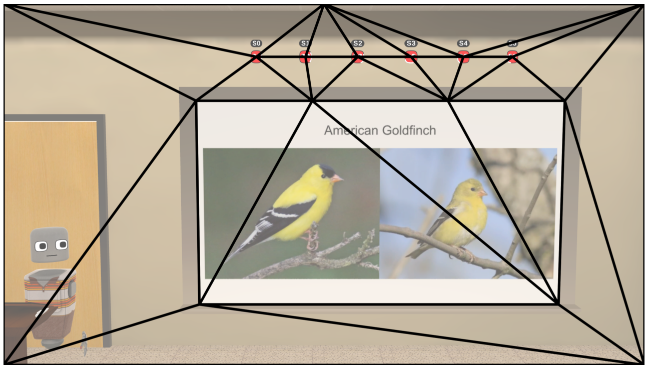

Figure 4: The teacher interface with overlayed Delaunay triangulation for the Mapped Gaze redirection, using key targets as triangle vertices. The triangulation is not visible during ordinary use. The interface shows a presentation board, a row of student icons above, and the teacher's avatar at a podium.

3.2 Teacher Interface

Cross-reality meeting spaces often give desktop users a first-person view rendered by a virtual camera placed at the head of a user-controlled character. This limits the scope of immediately interactable objects to those within camera view, constrained by monitor size and character pose (view direction). This makes it difficult to support certain natural interactions such as a teacher pointing to a board behind the teacher while looking out at students (a user would typically need to see the board to allow interaction with it).

Instead, our teacher interface, seen in Figure 3, renders the classroom from a central audience location, showing the presentation board and teacher avatar. This allows the teacher to focus on, and interact with, the presentation board without the distorted perspective of oblique viewing angles. This enables easy drawing on the board using mouse motion. Seeing the teacher avatar gives the teacher an understanding of how the avatar moves and responds to interactions. The teacher's avatar is placed at a podium with a tablet on top to mimic a common presentation setting.

A row of student icons is added to the teacher interface above the presentation board so that the teacher sees a representation of all students in a compact space without requiring view changes to face a student. The icons reveal student actions and attention such as eye gaze, hand raising, and chatting, using methods from [5]. Icons are dynamically added or removed when new students join or leave the system. More importantly for the redirection interface, the icons provide a way for the teacher to make eye contact with any student using the Mapped Gaze method (described below).

3.3 Redirection Methods

Redirection methods are modules in the redirection layer, and all can be active simultaneously. Only the modules making the highest-priority requests will have their animation requests fulfilled.

3.3.1 Gaze Redirection

We consider that a desktop gaze redirection system could allow the user to simply look at different areas or targets on the desktop screen to cause an avatar to gaze towards corresponding virtual objects. We use a desktop Tobii Eye Tracker to track the screen coordinate that the user looks at (mouse-based pointing can be substituted when eye tracking is not used).

Our *Mapped Gaze* redirection computes a target position in the 3D scene for the avatar to look at based on where the user looks on the desktop screen. First, potential gaze targets in the scene are identified. In our example, these include the head position of each student avatar (updated to reflect movement), the podium, and corners of the room. Each target corresponds to an icon or position on the desktop screen. For example, each student icon is mapped to an in-world student head position and the corners of the presentation board are mapped to the corners of a tablet on a podium. If the user looks at one of these desktop screen targets, the avatar will move its gaze towards the mapped position in the virtual room.

Gaze coordinates between defined screen targets are interpolated. For this, the screen elements are triangulated (Figure 4) using Delaunay triangulation [11] with each of the screen targets (reference points) acting as triangle vertices. In every frame, the system determines in which triangle the user's gaze point currently lies, calculates the barycentric coordinates of the gaze point within the triangle, and then applies those weights to the vertices' mapped world positions to find the appropriate gaze target position. The triangulation is recalculated if student icons are added or removed. As a result, the world gaze target is smoothly interpolated between points of interest as the user looks around the screen.

This method, when combined with the Body Mechanics system, presents more realistic human eye motions not possible with traditional desktop avatars. For example, when the teacher reads text on the presentation board, the avatar will gaze down at the podium, its pupils shifting side to side to reflect the teacher's actual reading eye motions. If a student needs to be addressed, a teacher can glance up towards the student's icon and the avatar's pupils will move up to make eye contact with the student with the subtle motions a teacher makes glancing up from a podium.

A limitation is that our implementation does not support gaze at unanticipated targets because targets must be predefined (gaze elsewhere is interpolated from them). We consider this reasonable for our virtual classroom. The implementation could be extended to consider more targets, automatically identify new or dynamic targets, or by casting view rays to find 3D target coordinates not near critical mapped regions.

3.3.2 Pointing Redirection

Pointing redirection enables the user's avatar to point at objects or positions in the environment. For example, the teacher can mouse over a desktop presentation board and left-click to direct their avatar to point at the matching position on the virtual board.

We discuss two methods: Grid and Algorithmic Pointing. Both methods determine a target hand pose and finger orientation to point a held laser pointer (with a long cylinder representing a laser ray). When not being controlled by hand tracking, finger rotations are set by interpolating from a set of pre-defined hand poses (e.g., gripping the pointer when pointing is initiated). Our stylized avatar lacks attached arms, avoiding the need for inverse kinematics for arm movement, and future work could add it for other avatars.

We consider that the avatar should often look at the board while pointing at it, like a human presenter. So, pointing redirection methods send the pointing target as a gaze target to the Priority System with a higher priority than standard gaze redirection. As a result, when a user triggers pointing, the avatar will turn around to look at the pointed-at position, and return when pointing concludes.

Grid Pointing Grid Pointing simulates human pointing by interpolating pose from a grid of recorded "keyframes" of actual human poses taken from an immersed VR-tracked actor. To capture keyframes, a 9×9 grid is textured onto the presentation board and the headset-immersed actor points with the laser pointer tool to each grid point, including corners and sides of the board. Actor head and hand poses are recorded at each grid point, resulting in 81 keyframe poses. To achieve reasonable smoothness in motions from interpolated poses, the actor points at the grid points one row at a time with slow smooth motions throughout.

When determining the avatar's target pointing pose, the system determines which grid cell the mouse cursor is contained in for bilinear interpolation from the four cell corners. Interpolated values include position and orientation of head and hands. To reduce erratic head movements, the target head position is only calculated when the user first initiates pointing. Pointing and moving the cursor across the board creates human-like hand movement, including slight inaccuracies and bobbing motions.

Slight hand rotation inaccuracies from human motion in recording initially led to the avatar's pointing ray appearing slightly off from the intended target. To solve this without fundamentally altering the hand poses, we apply the smallest possible hand rotation that aims the ray to precisely match the mouse input.

Algorithmic Pointing As an alternative to Grid Pointing, we added a method based on a simple kinematic algorithm. It uses the positions of the pointing target and the torso origin of the teacher avatar (roughly the upper center of the avatar's chest). The vector from the torso origin to the board target position is calculated, and the hand moved towards a fixed point along this vector, 0.6 meters from the torso, corresponding to a partly-extended imaginary arm. The hand's target rotation is re-calculated so that it points along the vector, then is rotated so that the thumb is above the rest of the hand.

Algorithmic Pointing avoids grid recording steps and avoids motion artifacts from extraneous changes in actor pose between keyframes. Due to the order in which keyframes are recorded, Grid movements look smoothest when performed along the primary direction or recording (horizontal movements in our case).

3.3.3 Drawing Redirection

Drawing on a board or presentation slides is an important aspect of lecturing and is supported by popular desktop presentation software. We allow a user to draw on the virtual board by right-clicking and moving the mouse on the desktop view. We present two redirection methods that produce the avatar drawing motion: Wand and Podium Drawing. Each method defines target poses for the drawing hand and sends the requests to the Priority System with a high priority.

Wand Drawing Wand drawing mimics an avatar behavior from immersive VR presentation with ray-based drawing. As the user draws on the board, it engages the selected pointing redirection method and changes the laser pointer's ray to show an ink color (red, in our study). The avatar gives the appearance that the teacher is drawing directly on the board with the laser. Drawings are rendered as polylines on the board and saved in a vector format for recall.

Podium Drawing In Podium drawing, the teacher's avatar moves a pen on a virtual tablet, at the teacher podium, to match what is being drawn on the desktop and board. The avatar behavior resembles the use of touch screens in "smart" classrooms.

A vector normal to the tablet surface is constructed using the four corners of the tablet. The drawing position is obtained from the board, in board-local space. A corresponding tablet coordinate for the pen tip is computed by interpolation from tablet corners. The drawing pen position is offset in the direction of the normal when the teacher stops drawing momentarily. If more than a second goes by without the teacher actively drawing, the method deactivates.

The hand and pen are rotated together slightly to simulate wrist motions associated with small drawing movements and hand orientation changes from elbow rotation. More specifically, a small rotation about the pen's axis is based on horizontal position to simulate hand/forearm angle from an (imagined) elbow. Horizontal and vertical drawing velocities control two rotations about tablet-parallel axes to simulate wrist motions and make the pen tip lead the hand.

3.3.4 Miscellaneous Redirection Modules

Other forms of redirection and animation are added to increase avatar liveliness and fill gaps in the user's motion when not engaging any of the previously discussed modules.

Idle Hand Animation As is common in desktop avatar systems, some minor animation is added to the hands to keep the avatar from looking stiff when speaking. Real human motion was captured from a VR user with tracked controllers in a 30-second clip to be replayed in a loop. The module constantly sends low-priority requests to the Priority System, causing other redirection methods to override it and returning to the idle animation afterwards.

Hand Tracking To support complex hand gestures, a Leap Motion Controller is included at the teacher's desktop. When the teacher moves a hand into the Leap's tracking bounds, the module will request the full hand configuration with the highest priority, overriding pointing or drawing when the teacher has a more specific gesture to show. As a result, the teacher can, for example, wave to students, make indications of size, or simply "speak with their hands."

Speech and Blinking Animation We animate the avatar mouth and blinking based on sensed user behavior. We did not include facial tracking, but it could be added from a webcam. The robot-like avatar does not support detailed facial expressions.

A stylized speaking cue is based on microphone input. Per frame, we adjust a waveform-like mouth drawing according to recent microphone samples. The waveform is the sum of two sine waves with amplitudes determined by average microphone input. Although it doesn't directly visualize the input audio waveform, this gives an abstract impression reminiscent of an oscilloscope, which is thematically consistent with the robot-like avatar. This mouth animation is used in all teacher conditions (described later), as most social VR spaces and remote meeting tools provide speaking indicators.

For blinking, when the eye tracker does not report a valid reading for a few milliseconds, we assume the teacher is blinking, and we close the avatar's eyes until a valid eye gaze is received. Our avatar's eyes are effectively large, flat images, and closing the eyes is done by scaling the eye image vertically (the pupil remains fixed to give the impression of animating an eyelid rather than the whole eye).

Blinks longer than 1 second are assumed to be due to technical problems. In this event, the blinking controller falls back to a canned randomized blinking animation, where the eye blinks once every 0.3 to 3.0 seconds for a short, random duration. The automated blinking is also used when there is no eye tracker active, and it is analogous to idle blinking animations in common social VR tools. Because the VR eye tracker reports eye openness as a value ranging from 0 to 1, we use this value directly to set the eye scale when using VR.

3.4 Priority System

The Priority System is the simplest layer. Each active redirection module sends a request to move specific body parts or the gaze to targets, along with a priority value. After all redirection request arrive, priorities are sorted and the highest-priority targets are set. Priority values are separate per body part and gaze target, allowing different redirection modules to control different body parts.

3.5 Body Mechanics

The Body Mechanics layer includes a PID controller [4] to move body parts towards target poses and a model of human eye and head motion to bring the eyes to gaze at a target. The outputs of the PID controller are also input to a torso rotation model and optionally combined with head tracking pose data.

Human motion, such as pointing, involves a rapid acceleration from start, a quick movement to the general target area, then rapid deceleration and small corrective motions to the target. Multiple methods for this style of movement were considered. For example, work from Yeo et al. [25] models submovements of real catching motion. We currently use a PID controller for its simplicity and ability to produce reasonable behavior with constantly moving target positions. Other methods that drive an object to a target pose could be plugged in to the body mechanics system in future extensions.

To animate head and eye rotations, we take inspiration from works describing natural gaze movement [7, 13]. For the gaze target (a world-space position) we consider two angles: the rotation required to align the pupils to face the gaze target (termed gaze angle), and the angular offset between the pupils' current positions and their eye's center (termed pupil angle).

The pupil is rotated at 500° per second to minimize the gaze angle. If the pupil angle is less than 20°, the head does not rotate and the pupils make regular saccadic motions. If the pupil angle is between 20° and 30°, the pupil angle is brought back below the 20° threshold by making a slow head rotation at a rate of 30° per second. This is done to mimic the human tendency to minimize eye strain by keeping the pupils from straying too far from the eye center. If the pupil angle exceeds 30°, we assume the new target would be out of the avatar's view, and make a fast head rotation of 300° per second to bring the new target into view. As a result, the avatar is able to make only subtle pupil and head movements when reading words from the podium or looking between students, but quickly turns around to look at the board when initiating pointing.

3.5.1 Supplemental Movement

Torso Movement Initial versions of the redirected avatar simply placed the torso a set distance down from the avatar head and rotated it based on a threshold as the head rotated. Resulting motion may appear stiff and unrealistic. To address this, we consider that torso movement is affected by hands as well as head [10] and thus include hands in the rotation calculation. A torso-attached coordinate frame controls hand rotation around the body by following the rotation of the head with a 45° threshold to trigger rotation. The torso mesh separately orients to face the midpoint between the two hands. As a result, the torso will naturally produce small rotations as the user makes subtle hand motions, and will make wider swings as the user, for example, looks from the board to the audience.

Supplemental Head Motion Because a standard desktop does not track head motion, typical desktop-based avatars have minimal or canned head motion compared to head-tracked VR. We calibrate the user's resting head pose at the beginning of the session. The system then continuously reads the user's head position and orientation from the Tobii Eye Tracker, subtracts that from the resting head pose in the center of the user work space, and moves the avatar's head by the resultant amount. As a result, the avatar's head is livelier and allows the teacher to make gestures such as nodding.

4 User Study

The user study compares our novel desktop teacher avatar, driven by the various sensors described above, to avatars animated by a tracked VR headset user, to a more conventional first-person desktop controller, and to our desktop interface without the use of the Tobii Eye Tracker or Leap Motion Controller. Our intent was to see if users have a preference in the method of animation, if that preference would translate into behavioral response, and if people would view our avatar as human-like or strange when compared to the more conventional means of animation.

Thirty-six subjects (33 male and 3 female, aged 20 to 33, mean age 22.6) were recruited from a local computer science department. Subjects took the role of a student viewing prerecorded behaviors of the avatar produced by the four animation conditions detailed below. Playback of recordings controlled motion and audio aspects of the teacher avatar, which was placed in the virtual classroom wherein the subject had a standard immersive VR view (tracked head and hands). Subjects were given limited opportunity to interact except for answering prompts.

There were two study phases. In Phase 1, subjects watched and gave opinions about four approximately two-minute presentations by the four teacher avatars presenting facts about local birds. In Phase 2, subjects watched short animations showcasing a single type of redirected motion and gave opinions and rankings.

4.1 Independent Variable

The independent variable was the type of teacher avatar animation control, with the four conditions below:

- *VR* - Conventional immersive VR control with the teacher animated by tracking of a Vive Pro Eye headset and controllers. The result is similar to the VR interface in Mozilla Hubs, but we add eye tracking to animate pupils and blinking.
- *CDesk* - Conventional first-person desktop control, modeled primarily after Mozilla Hubs. The user's primary controls include clicking the middle mouse button and dragging the mouse to rotate the avatar head and camera. To produce similar behaviors to the other conditions, when the user is facing the board, they can activate drawing or pointing with the mouse as described in the redirection system above. This uses the algorithmic pointing method and the wand drawing method. It does not use the Body Mechanics system, instead moving body parts to targets with a constant velocity. Blinking is randomized and pupils remain centered.
- *SDesk* - Desktop with full sensing. The teacher is animated by our methods described above, including eye, head, and hand tracking. We are primarily interested in seeing the effects of this avatar compared to the VR and CDesk.
- *NDesk* - Our teacher interface with no special sensors. This omits hand tracking and Supplemental Head Motion modules. Mapped Gaze uses the mouse position instead of an eye tracker. Our intent is to discern the impact of the additional sensors (and the redirection they enable) on our novel interface, and see what is possible with more ordinary desktop setups.

Sixteen avatar motion recordings were made by a single teacher experienced with each interface. The 16 recordings covered each combination of the 4 conditions and 4 birds about which a presentation was given. This allowed each subject to view 4 different bird presentations to avoid repetition while supporting the randomization described below. For consistency, a single audio recording was used per bird presentation, with a different avatar motion recorded for each avatar condition. During recording, the teacher could look at a certain student icon for which gaze would be redirected to the subject during trials.

Across all recordings, the teacher made similar presentation motions and maintained similar amounts of eye contact to the extent allowed by the different conditions. To that end, each presentation had 4-5 slides covering the same type of content. Pointing was performed on each slide and drawing performed on two, with eye contact made at natural intervals between them.

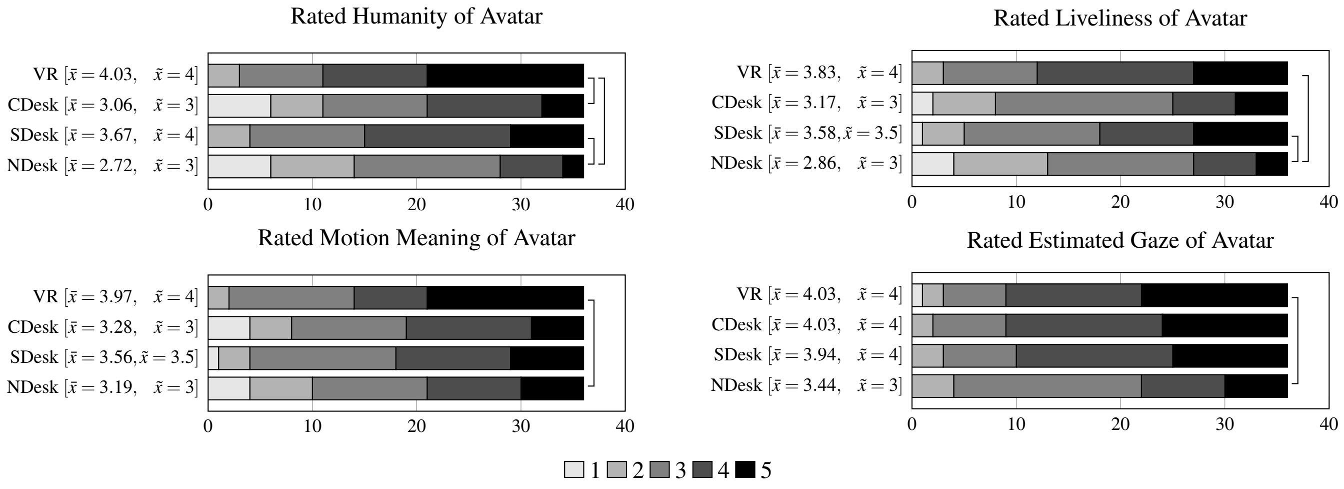

Figure 5: Subject ratings from student perspective of teacher avatar for questions asked in Phase 1, including mean and median listed for each condition. Brackets indicate statistically significant differences following Bonferroni correction. Results from the quiz and co-presence question are excluded because no significant difference was shown for them.

Each subject experienced four avatar-bird combinations such that each condition was experienced once and each bird was described once. The four combinations appeared in a random order. Across subjects, the total number of times each combination appeared to any subject was counted. Per subject, a randomizer selected from the least-used combinations to ensure all combinations were used approximately the same number of times across the subject pool. So, all 16 recordings were played 9 times total across 36 subjects.

4.2 Dependent Variables

Our study is primarily concerned with the perceived humanity of the teacher's avatar, the engagement, and preference for redirection methods. In Phase 1, subjects were presented with the following six questions on an in-world panel (answered by ray pointing):

1. *Quiz* - A multiple choice item about the presented bird.
2. *Humanity* - "Were the teacher's motions robotic or human-like?"
3. *Lively* - "How lively was the teacher?"
4. *Co-Presence* - "How much did it feel like the teacher was in the room with you?"
5. *Motion Meaning* - "How meaningful were the teacher's motions?"
6. *Estimated Gaze* - "How often was the teacher looking at you?"

Quiz questions were asked to encourage subjects to pay attention to content and were not considered a main item for analysis. Questions 2-6 were asked as 5-point Likert-like items.

In Phase 2, subjects ranked the four avatar conditions for pointing, drawing, and greeting motion types, based on preference and motion strangeness. They then answered which interface they think each avatar used from the options "A VR headset and controller," "Interaction devices at a desktop," and "A motion-generating AI." They also gave their preference between the two pointing and drawing redirection methods.

4.3 Study Procedure

After signing consent forms and completing background questionnaires, subjects donned a Vive Pro Eye headset. The eye tracking was calibrated per user. The subject was given a brief experiment overview by a recorded static avatar showing only mouth motion.

In Phase 1, 4 avatar recordings were chosen by the randomizer as described before. After viewing each recording, the subjects answered the Phase 1 questionnaire. After all recordings, the subject was verbally asked to describe what factors they used to determine if a teacher was robotic or human-like.

Phase 2 had three stages. In the first stage, three sets of four 10-second recordings of each avatar were played. Each set showed pointing, drawing, or greeting behavior. Pointing and drawing motions made simple gestures towards images on the presentation board. The greeting motion looked up from the podium, addressed the subject, and waved if possible. After each set, the subject ranked the four recordings by preference, then by strangeness, with the ability to replay each recording on demand. The order of sets and the order of recordings within sets was random. After all three sets, the subject was asked what motion factors they considered for ranking.

In the second stage, four recordings were played, in random order, of each avatar condition for a drawing motion. The subject answered what interface they believed the teacher was recorded with.

In the third stage, two sets of two recordings of SDesk pointing and drawing motions were played. Each recording used one of either Grid or Algorithmic Pointing redirection, and one of either Wand or Podium Drawing redirection, respectively. After each set, the subject picked which motion they preferred. Again, the order of sets and the order of recordings within them were randomized.

5 Results and Discussion

5.1 Phase 1 Questionnaires

Phase I ratings are summarized by Figure 5 per question and avatar condition. Ratings were analyzed by Friedman tests with Bonferroni-corrected Wilcoxon signed-rank followups (reported p-values include multiplication by 6 for Bonferroni correction).

Quiz answers were encoded as correct or incorrect based on the subject's answer. A related-samples Cochran's Q test did not show significant differences between avatar conditions ($\chi^2(3) = .343, p = .952$). This is not surprising because quizzes were intended only to encourage student attention, and we did not expect to measure significant educational differences in these brief presentations.

The *humanity* ratings show a difference between avatar conditions based on a Friedman test ($\chi^2(3) = 29.810, p < .001$). Post-hoc tests show VR rated more human-like than NDesk ($p < .001$) and CDesk ($p = .024$), and SDesk rated more human-like than NDesk

($p = .004$). Assuming that users will prefer to inhabit a virtual world with an avatar they consider more human than robot (as was implied by interviews with subjects after the experiment, and is supported by later results), this implies that the NDesk and CDesk avatars are less preferable to what may be considered the “gold standard” VR avatar. The SDesk avatar was not detected different from the VR avatar ($p = .438$). While this does not mean that subjects considered them identical, it would be consistent with a less notable difference than from NDesk and CDesk.

The *lively* ratings also differ ($\chi^2(3) = 19.289, p < .001$), with VR and SDesk both ranking higher than NDesk ($p = .004$ and $p = .042$ respectively). Here, there is a near consistency between the perceived humanity of the avatar and its liveliness. Indeed, when interviewed about what traits subjects looked for when determining humanity, 28 stated their opinions were based on some variation of frequent, smooth, and lively movement.

The *motion meaning* ratings differ ($\chi^2(3) = 14.525, p = .002$), with VR ranking higher than NDesk ($p = .016$). Readers may also consider a likely trend of VR rating higher than CDesk ($p = .064$) considering that Bonferroni correction tends to be overly conservative. *Gaze estimation* ratings differ ($\chi^2(3) = 11.822, p = .008$), with VR again ranking higher than NDesk ($p = .037$).

The *co-presence* rating (not plotted) shows a difference overall ($\chi^2(3) = 10.579, p = .014$) but post-hoc tests do not identify a specific difference after Bonferroni correction (VR vs. NDesk hinged on Bonferroni correction, with pre-corrected $p = .02$).

When asked what contributed to a high *humanity* rating, most subject responses (28) reflected fluidity of motion or slower, controlled speed. This is in line with results, as the CDesk controls make it difficult to consistently produce smooth, fluid motion, compared to the VR controls. This also benefits NDesk and SDesk approaches as the values in the Body Mechanics layer can be adjusted for smoother motion in a way not possible for CDesk. The significant difference between NDesk and SDesk may be explained by the motions from extra sensors, such as subtle head bobs, faster saccades, and natural hand gestures. This is reinforced by 11 subjects stating they liked when the teacher waved. When asked if they preferred an avatar that was more human-like, all subjects said yes.

Notably, although VR was found significantly better than CDesk (and NDesk), SDesk was not shown different from VR or CDesk by statistical tests. This indicates SDesk differs from at least one of CDesk and VR, but does not confirm which. We believe the most likely explanation is that tradeoffs place SDesk somewhere between VR and CDesk. Humanity and Liveliness means suggest that SDesk is more likely to differ from CDesk, i.e., that SDesk is more likely close to VR. Additionally, SDesk is found significantly better than NDesk, implying the addition of sensors improved the avatar. NDesk could potentially benefit from adding, e.g., automated head movements or hand gestures to mimic the SDesk avatar.

5.2 Phase 1 Eye Metrics

During playback of Phase 1 recordings, we recorded the angle between the subject’s gaze direction and the direction from the subject’s eye center to the teacher’s eye center, at 120 Hz. We compared summary values (mean, median, and standard deviation) for the angle per condition between subjects.

A significant difference between conditions was found in mean angle by the Friedman test ($\chi^2(3) = 12.6, p = .006$) with pairwise tests showing a higher average angle for CDesk compared to VR ($p = .02$) and SDesk ($p = .027$). A lower average angle for VR and SDesk suggests that subjects spent more time looking towards those avatars than the CDesk version. We suspect this is due to their livelier and more human-like natures. No difference was detected in median or standard deviation.

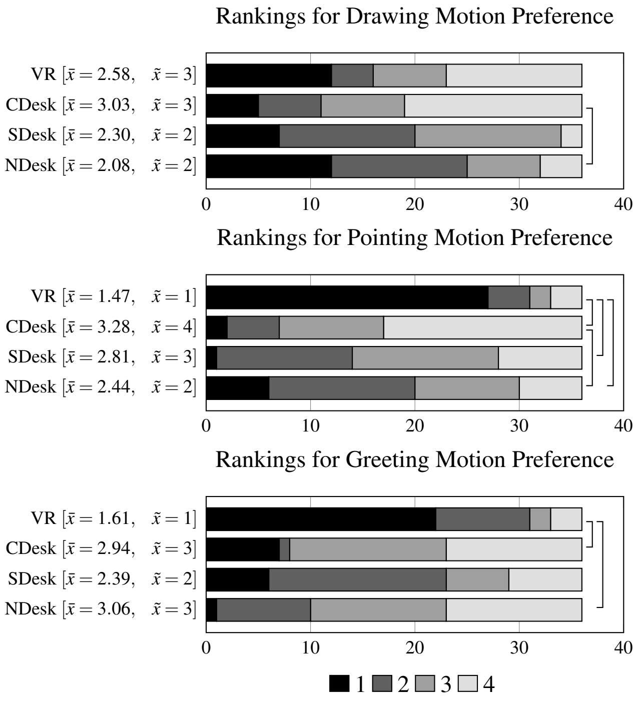

Figure 6: Preference ranking results for three types of motions compared between avatars, including mean and median rank. A lower number indicates a more favorable ranking. Brackets indicate statistically significant differences following Bonferroni correction.

5.3 Phase 2 Results

5.3.1 Ranking Stage

Ranking by subjects gave a 1 to the highest ranked condition (1st) and a 4 to the lowest (4th). A rank of 1 for a preference question indicated the most preferred condition, and rank 1 for a strangeness question indicated the strangest condition. Rankings for preference are summarized in Figure 6.

There was a significant difference in drawing motion preference ranks ($\chi^2(3) = 10.733, p = .013$). Post-hoc tests showed NDesk was ranked better than CDesk ($p = .011$). A Friedman test for drawing motion strangeness showed no differences between conditions ($\chi^2(3) = 4.1, p = .251$). Subject interviews revealed divergent opinions about preferred drawing motion. As a natural result of VR wand pointing and desktop mouse input methods, drawings in the VR condition appeared uneven and shaky while those drawn in desktop conditions were smoother. This was particularly pronounced when drawing circles. Some users noted a preference for the shakier drawing to match the shakiness of real human pointing, while others preferred the smoother lines because they were more visually appealing, possibly explaining the lack of a clear ranking.

There was also a difference in pointing motion preference ranks ($\chi^2(3) = 37.967, p < .001$), with post-hoc tests showing VR ranked higher than NDesk ($p = .008$), SDesk ($p < .001$), and CDesk ($p < .001$). NDesk ranked higher than CDesk ($p = .037$). Pointing strangeness ranks also showed differences ($\chi^2(3) = 20.1, p < .001$) with CDesk ranked stranger than VR ($p < .001$) and SDesk ranked stranger than VR ($p = .002$). Subject interviews revealed that the high VR rankings were likely because of the slow, smooth motion of its hand to the pointing position.

Greeting motion ranks showed differences ($\chi^2(3) = 28.267, p < .001$), with VR ranked higher than CDesk and NDesk ($p < .001$ for both). Greeting strangeness ranks showed differences ($\chi^2(3) =$

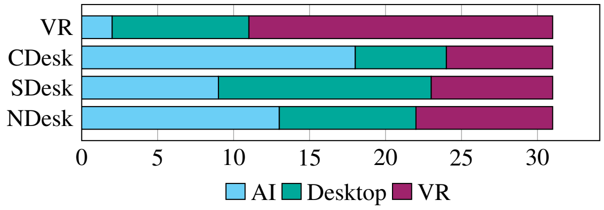

Figure 7: Results from Phase 2's question asking subjects which type of interface or motion source they thought each condition used. Results are missing for 5 of the 36 subjects due to technical problems affecting only this stage.

$23.633, p < .001$), with VR ranked less strange than NDesk ($p < .001$) and CDesk ($p = .028$). VR had significantly better ranks than other conditions, likely because of the ability to easily wave and make eye contact, suggesting such gestures can be valuable. SDesk is not shown to differ from other conditions; after inspection of averages we consider the most likely explanation to be that SDesk falls somewhere in between VR and CDesk, suggesting a need to improve its greeting motions to bring it closer to VR's rankings.

5.3.2 Perceived Interface Stage

Answers from the perceived interface question are summarized in Figure 7. We consider it valuable if the subject thinks the avatar is animated by a user in VR or at a desktop, because it could increase the chance of perceiving human teacher presence during live use, in contrast to AI-style motion. In line with findings from Phase 1 on the humanity of avatars, it appears that most subjects (29) consider the VR condition to be human-controlled, followed by the SDesk and NDesk conditions (22 and 18 subjects, respectively). Most subjects indicated that CDesk looked AI-controlled.

5.3.3 Preference Stage

When choosing between Wand and Podium Drawing, 23 subjects preferred Wand and 11 chose Podium. When asked to explain their choice, those who chose Wand stated that its simplicity kept their attention on the board where information was being presented, instead of dividing their focus between the board and the teacher avatar. They also liked that the interaction was something that the virtual classroom enables that is not possible in the real world. Those who chose Podium stated that they liked that the motion felt more realistic and like something they had seen in smart classrooms.

The preference between Grid and Algorithmic Pointing is less clear, with 16 subjects choosing Algorithmic and 18 choosing Grid. Those who chose Algorithmic stated that they liked that the hand moved closer to the board, while those who chose Grid stated they felt the orientation of the Algorithmic hand and its implied straight-out arm was strange. Those who chose Grid also stated they felt it was a more realistic and natural motion for the use of the laser pointer tool. This result suggests some pose tuning could adjust each approach to reduce differences or consider user preference.

6 Conclusion and Future Works

This work designed and tested methods for animating a cross-reality avatar using a novel desktop interface to more completely reflect a desktop presenter's intended motions to a VR audience. With a three-tier redirection architecture, we provided lively avatar motion capable of most gestures needed for presentations in VR. Secondary goals included seeing which of our redirection methods were preferred, determining the value of the additional sensors in our interface, determining if users did actually prefer a VR avatar to a conventional desktop avatar, and seeing if our sensor-based desktop avatar could achieve similar ratings to a VR avatar.

While the headset-controlled avatar was rated highest in all aspects except drawing motion, a redirected desktop avatar was almost always second in rank. Phase 2 of the study showed that pointing and drawing motions were better received when performed with redirected methods (NDesk) when compared to conventional methods (CDesk). Subject interviews revealed a preference for slow, smooth, and fluid avatar motion. While a headset-controlled avatar is most capable of achieving this, the perspective of our desktop interface allows for more fluid pointing and drawing motion than conventional avatars that involve stiff motion and perspective changes to engage pointing. Though subjects thought our desktop avatars moved too fast, leaving a negative impact on their humanity rankings, the parameterized nature of the body mechanics system would allow this to be tuned more easily than a conventional avatar.

When used without added sensors (hand and eye trackers), our desktop avatar was rated less human-like and less lively than the VR avatar, but still ranked highly among individual motion preferences. This suggests that our outlined techniques may still be useful in settings where no additional sensors are available.

In several comparisons, the conventional desktop avatar was rated worse than the VR avatar. This is expected, and shows the difference in quality that users of current educational cross-reality applications may be receiving when viewing desktop-controlled avatars. Though some tradeoffs exist when compared to VR, the fact that most users considered the conventional desktop avatar to appear AI-controlled, the increased attention to the SDesk avatar's head, and the likelihood for SDesk to be closer in humanity to the VR avatar suggest that SDesk provides an improved desktop-controlled avatar.

The usability of our interface was not evaluated from a teacher's perspective, because we focused on the student experience first. Future work could evaluate the usability of different interface designs for the cross-reality presentation environment, for example, changing the teacher's view to omit their own avatar, changing the layout and representation of student icons, and changing input methods for drawing or pointing. Additional redirection methods could also be added to the interface as non-invasive sensors become more prominent and image processing methods become more robust, for example, webcam-based tracking of eyes or facial expressions. Considering the short exposure times in Phase 2, future studies on a student's perspective could investigate longer lecture times or more recordings of individual motions.

Other teacher avatar types could be considered, for example, to see if the methods work with human-like avatars. Some user feedback on avatar motion could be addressed, tuning movement parameters to slow down and smooth out the motions. Other controllers could be tried in place of the Body Mechanics PID controller to further model human motion aspects. Additional gaze behaviors such as transformed gaze [3] or occasional automated glances at each student may improve the student experience. Given this work's focus on the specialized, static teacher perspective, similar redirection techniques could be applied or modified to work in other meeting environments or dynamic perspectives.

Acknowledgments

This paper is based on work supported by the National Science Foundation under Grant No. 1815976 and by the Louisiana Board of Regents Support Fund grant LEQSF(2019-20)-ENH-DE-22.

References

[1] S. J. G. Ahn, L. Levy, A. Eden, A. S. Won, B. MacIntyre, and K. Johnsen. Ieeevr2020: Exploring the first steps toward standalone virtual conferences. *Frontiers in Virtual Reality*, 2:28, 2021. doi: 10.3389/frvir.2021.648575

[2] S. Andrist, M. Gleicher, and B. Mutlu. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 2571–2582. Association

for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3026033

[3] J. N. Bailenson, A. C. Beall, J. Loomis, J. Blascovich, and M. Turk. Transformed social interaction, augmented gaze, and social influence in immersive virtual environments. *Human communication research*, 31(4):511–537, 2005.

[4] S. Bennett. Development of the pid controller. *IEEE Control Systems Magazine*, 13(6):58–62, 1993. doi: 10.1109/37.248006

[5] D. M. Broussard, Y. Rahman, A. K. Kulshreshth, and C. W. Borst. An interface for enhanced teacher awareness of student actions and attention in a vr classroom. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 284–290, 2021. doi: 10.1109/VRW52623.2021.00058

[6] P. Engzell, A. Frey, and M. D. Verhagen. Learning loss due to school closures during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 118(17), 2021. doi: 10.1073/pnas.2022376118

[7] E. G. Freedman. Coordination of the eyes and head during visual orienting. *Experimental brain research*, 190(4):369–387, 2008.

[8] D. Garrison, M. Cleveland-Innes, and T. S. Fung. Exploring causal relationships among teaching, cognitive and social presence: Student perceptions of the community of inquiry framework. *The Internet and Higher Education*, 13(1):31–36, 2010. Special Issue on the Community of Inquiry Framework: Ten Years Later. doi: 10.1016/j.iheduc.2009.10.002

[9] T. Kim, A. Kachhara, and B. MacIntyre. Redirected head gaze to support ar meetings distributed over heterogeneous environments. In *2016 IEEE Virtual Reality (VR)*, pp. 207–208, 2016. doi: 10.1109/VR.2016.7504726

[10] S. Kita. Interplay of gaze, hand, torso orientation, and language in pointing. In S. Kita, ed., *Pointing*, pp. 307–328. Psychology Press, 1st ed., 2003.

[11] D.-T. Lee and B. J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.

[12] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[13] D. E. Mitchell and K. E. Cullen. Vestibular system. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier, 2017. doi: 10.1016/B978-0-12-809324-5.02888-1

[14] P. Pazhayedath, P. Belchior, R. Prates, F. Silveira, D. S. Lopes, R. Cools, A. Esteves, and A. L. Simeone. Exploring bi-directional pinpointing techniques for cross-reality collaboration. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 264–270, 2021. doi: 10.1109/VRW52623.2021.00055

[15] T. Piumsomboon, G. A. Lee, J. D. Hart, B. Ens, R. W. Lindeman, B. H. Thomas, and M. Billinghurst. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018.

[16] T. Randhavane, A. Bera, K. Kapsaskis, K. Gray, and D. Manocha. Fva: Modeling perceived friendliness of virtual agents using movement characteristics. *IEEE transactions on visualization and computer graphics*, 25(11):3135–3145, 2019.

[17] M. Rebol, C. Gütl, and K. Pietroszek. Real-time gesture animation generation from speech for virtual human interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2021.

[18] A. Riegler, C. Anthes, H.-C. Jetter, C. Heinzl, C. Holzmann, H. Jodlbauer, M. Brunner, S. Auer, J. Friedl, B. Fröhler, et al. Cross-virtuality visualization, interaction and collaboration. In *XR@ ISS*, 2020.

[19] D. Roth, C. Klelnbeck, T. Feigl, C. Mutschler, and M. E. Latoschik. Beyond replication: Augmenting social behaviors in multi-user virtual realities. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 215–222. IEEE, 2018.

[20] D. Roth, P. Kullmann, G. Bente, D. Gall, and M. E. Latoschik. Effects of hybrid and synthetic social gaze in avatar-mediated interactions. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 103–108. IEEE, 2018.

[21] A. L. Simeone, M. Khamis, A. Esteves, F. Daiber, M. Kljun, K. C. Pucihar, P. Isokoski, and J. Gugenheimer. International workshop on cross-reality (xr) interaction. *Companion Proceedings of the 2020 Conference on Interactive Surfaces and Spaces*, 2020.

[22] A. L. Simeone, M. Speicher, A. Molnar, A. Wilde, and F. Daiber. Live: The human role in learning in immersive virtual environments. In *Symposium on Spatial User Interaction*, pp. 1–11, 2019.

[23] R. Vertegaal, I. Weevers, and C. Sohn. Gaze-2: An attentive video conferencing system. In *CHI'02 extended abstracts on Human factors in computing systems*, pp. 736–737, 2002.

[24] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction. *Frontiers in Virtual Reality*, 2, 2021. doi: 10.3389/frvir.2021.641296

[25] S. H. Yeo, M. Lesmana, D. R. Neog, and D. K. Pai. Eyecatch: Simulating visuomotor coordination for object interception. *ACM Transactions on Graphics*, 31, 2012. doi: 10.1145/2185520.2185538

[26] A. Yoshimura and C. W. Borst. A study of class meetings in vr: Student experiences of attending lectures and of giving a project presentation. *Frontiers in Virtual Reality*, 2:34, 2021. doi: 10.3389/frvir.2021.648619