An Exploratory Study of Augmented Reality Presence for Tutoring Machine Tasks

Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, Karthik Ramani
School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907 USA
{cao158, qian85, wang3259, lee2422, khuo, ramani}@purdue.edu

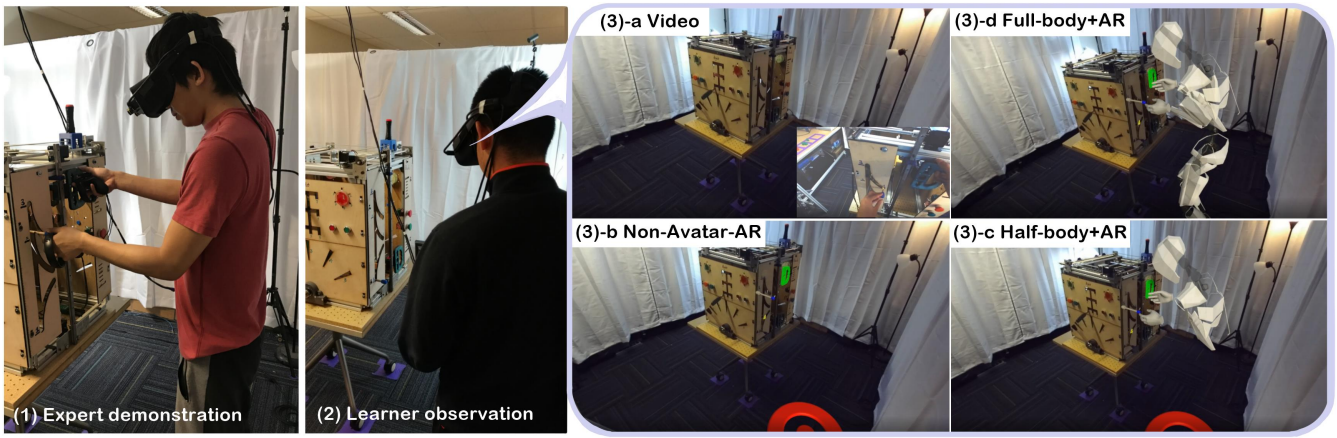

Figure 1. An overview of our exploratory study setup. An expert first generates a tutorial of a machine task on the mockup machine through embodied demonstration (1). Later a student tries to repeat the task by following this tutorial through an augmented reality (AR) headset (2). We propose to explore four tutor presence options for machine task tutoring, including: *video* (3)-a, *non-avatar-AR* (3)-b, *half-body+AR* (3)-c and *full-body+AR* (3)-d.

ABSTRACT

Machine tasks in workshops or factories are often a compound sequence of ***local***, ***spatial***, and ***body-coordinated*** human-machine interactions. Prior works have shown the merits of video-based and augmented reality (AR) tutoring systems for *local* tasks. However, due to the lack of a bodily representation of the tutor, they are not as effective for *spatial* and *body-coordinated* interactions. We propose avatars as an additional tutor representation to the existing AR instructions. In order to understand the design space of tutoring presence for machine tasks, we conduct a comparative study with 32 users. We aim to explore the strengths/limitations of the following four tutor options: ***video***, ***non-avatar-AR***, ***half-body+AR***, and ***full-body+AR***. The results show that users prefer the *half-body+AR* overall, especially for the *spatial* interactions. They have a preference for the *full-body+AR* for the *body-coordinated* interactions and the *non-avatar-AR* for the *local* interactions. We further discuss and summarize design recommendations and insights for future machine task tutoring systems.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

 ISBN 978-1-4503-6708-0/20/04…$15.00
DOI: https://doi.org/10.1145/3313831.3376688

Author Keywords

Machine Task; Avatar Tutor; Tutoring System Design; Exploratory Study; Augmented Reality.

1 INTRODUCTION

Contemporary manufacturing facilities are changing to focus on flexible, modular, and self-configuring production, a trend that is sometimes called *Industry 4.0* [40]. Human workers, as the most adaptive part of the production process, are expected to operate various machinery and equipment in a constantly changing working environment [22]. This creates a new challenge that requires workers to rapidly master new machine operations and processes, what we refer to in this paper as ***machine tasks***. Researchers have proposed low-cost, easy-to-distribute, and highly-scalable machine task tutoring systems as a way to resolve this challenge. Recent novel tutoring systems show potential to reduce and eventually eliminate real-human one-on-one tutoring [21].

Machine tasks in a workshop or factory environment are usually a mixed sequence of various types of interactive steps. Based on our observations and literature reviews, we categorize the steps of the machine tasks into three types: ***local***, ***spatial***, and ***body-coordinated*** [58, 32]. A *local* step refers to one-hand interactions in the user's immediate vicinity (i.e., within arms reach), which involves no spatial movement. A *spatial* step requires a large spatial navigation before proceeding to interact with the target machine interface. And in a

body-coordinated step, an operator must coordinates his/her body, hands, and eyes to complete the interaction.

Video content has been widely adopted into modern tutoring systems because they are capable of illustrating the fine details of operations [61, 23, 10, 9, 34]. Despite their popularity, video tutorials fundamentally suffer from the lack of a spatial connection between the digital representation and the user's physical presence. This flaw of video tutorials can lead to a fractured learning experience, especially for physically interactive tasks. To address this challenge, augmented reality (AR) approaches have been proposed that superimpose virtual tutorial guidance directly onto the interaction target in-situ [15]. Due to this advantage, AR tutoring systems have been particularly favored for interactive tasks within the physical environments, such as in machine-related operations [55, 30, 41, 66].

However, existing AR tutoring systems for machine-related operations predominantly focus on *local* interactions. The virtual tutoring contents in these works usually apply visual illustrations, such as static and dynamic symbols and text, to represent the operations within the local regions of interest. Previous works have shown their effectiveness for highly-complex local instructions, such as computer assembly [66], machinery diagnosis [67], and vehicle maintenance [4]. However, due to the lack of an explicit visual representation of the human tutor for spatial and bodily movements, these symbol-only AR illustrations are inadequate to provide clear cues for interactions that require large spatial navigation movements and full body coordinative operations, such as the machine tasks.

To guide the development of improved AR tutoring systems, we propose to use avatars as an enhanced tutoring presence to the existing AR instructions. In our approach, the embodied demonstration of the tutor is presented in the operator's AR view while they interact with the physical machines in-situ. Virtual avatars have been broadly used to represent the embodiment of the human users in various virtual reality (VR) consumer applications, such as VR-chat [64]. Avatars have also been explored and adopted in the area of mixed reality (MR) remote assistance [48, 51, 49], body movement guidance and training [68, 11, 3], and telepresence AR conference [46, 18]. Most recently, Loki [60] has demonstrated the avatar's potential for facilitating physical tasks via remote instructions. However, a systematic study of an avatar based AR presence is still lacking, especially in the context of machine task tutoring.

To this end, we investigate two research questions to reveal future research directions for the design of machine task tutoring systems. (i) Is the additional avatar presence beneficial to the user's experience and performance in a comprehensive machine task tutoring scenario, compared with the *non-avatar-AR* and *video* tutorial options? (ii) How to optimize the design of the tutor presence to achieve improved tutoring experience for future machine task applications?

To answer these questions, we develop two different avatar tutor presentations: *half-body* and *full-body*. Together we compare the following four tutor presence options: ***video***, ***non-avatar-AR***, ***half-body+AR***, ***full-body+AR***. Along these options, we gradually increase the guidance visualization levels, aiming to provide insights for an ideal design. All four options of the machine task tutorials are created from one single source, which is the embodied physical demonstration of the expert human tutor, as illustrated in Figure 1. We conduct a study with 32 users across four different tutor options, with a specially created mockup machine as the machine task testbed. The contributions of our paper are as follows.

- **Study System Design and Implementation** of a machine task scenario to compare all four tutor options in parallel, where *local*, *spatial*, and *body-coordinated* interactions are composed into multi-step tutorial sessions.
- **Quantitative and Qualitative Results** showing users' objective/subjective responses and tutor preferences after completing the sessions of machine tasks while following different tutor options.
- **Design Recommendations and Insights** summarized from the results and discussions of the study, providing valuable guidance for future machine task tutoring system design.

2 RELATED WORK

2.1 AR Tutorials for Machine-related Operation

AR naturally supports spatially and contextually aware instructions for interacting with the physical environment. Researchers have explored various designs for AR-based text instructions [67, 4, 70, 47], including numerical values [55, 8] for precise operational descriptions with quantitative real-time feedback. Symbolic visual guidance, such as arrows [20, 30], pointers [36], circles [33], and boxes [66], are commonly used for visualizing motion intent and guiding a user's attention. Besides text and symbols, prior works have also explored virtual 3D models of the interactive tools and machine components for a more comprehensive and intuitive visual representation, in use cases such as object manipulations and geometric orienting operations [41, 71, 65, 44].

These means for creating AR instructions have been useful for tutoring physical tasks. AR-based training systems have been thoroughly explored and applied to complex real-world scenarios, such as vehicle maintenance training [4, 14, 33], facility monitoring [70, 71, 67], machine tool operations [41, 8, 55], and mechanical parts assembly [30, 66, 65, 47]. However, most of these AR-based training systems are focused on *local* interactions that involve very little spatial navigation and bodily movement as a part of the human-machine task itself. To incorporate human motion into the task instruction, we propose virtual avatars for externalizing the human tutor. We acknowledge the necessity of the AR instructions in the existing work and additionally propose an avatar as a supplementary tutoring presence, mainly for the *spatial* and *body-coordinated* interactions in the machine task scenarios. We are interested in finding out if the added avatar visualization would improve the users' machine task tutoring experience and provide additional benefits that will inspire the future designs of intelligent tutoring systems.

2.2 Virtual Humanoid Avatar in AR/VR Training Systems

A virtual humanoid avatar is an animated human-like 3D model that embodies the human user's body movements, gestures, and voice information in VR and AR environments. It has been adopted as an expressive visualization media for human motion training. Chua et al. [12] built a Tai Chi training platform with a virtual instructor performing prerecorded movements, where the students follow and learn asynchronously. YouMove [3] utilized an AR mirror to achieve full body gesture comparison with a projected tutor avatar. In terms of providing a better comparison with the virtual instructor, previous works [26, 31, 28] superimposed the virtual instructor together with the user's perspective in the AR view, enabling the user to align his/her body spatially with the virtual avatar. Moreover, OutsideME [68] adopted virtual avatars to externalize the users themselves as a real-time reference so that they can see their own bodies from a third-person view while dancing. While differentiating from regular-sized avatars, Piumsomboon et al. [48, 49] exploited a miniature avatar to empower collaboration between a local AR user and a remote VR user. Most recently, Loki [60] has created a bi-directional mixed-reality telepresence system for teaching physical tasks by facilitating both live and recorded remote instructions via avatars and RGBD point cloud.

These previous works reveal the virtual avatar's advantages in enhancing bodily-expressive human-human communication, for applications such as asynchronous learning, self-observing and training, teleconference, and MR remote collaboration. Nevertheless, the usability of the avatar as a tutor presence for training in physically interactive tasks has not been systematically explored. This paper proposes to use avatars for representing the human tutor's spatial and bodily movements in the machine task training scenario. A machine task is a compound mixture of multiple types of interaction, and existing tutorial visualizations do have their own advantages. Therefore, it is paramount for us to study *when* and *how* to use avatars in order to apply it effectively in machine task tutoring.

2.3 Authoring by Embodied Demonstration

An embodied demonstration enables a user to use the shape, positioning, and kinematics of one's body as spatial reference for digital content creation. Researchers have achieved complex hand-related 3D sketching [37], design of personalized furniture [38], and creative 3D modeling [69]. Additionally, the motion data of the demonstrations can be extracted from videos to produce step-by-step training tutorials for human body action [3, 11], first-aid procedure [17], and parts assembly [25]. Similarly, by mapping extracted body motion to virtual characters, users can act out stories and generate animations directly [6, 29, 24, 52]. The embodied demonstration has also been applied in the area of human-robot interaction. Vogt et al. [63] and Amor et al. [2] used motion data captured from human-human demonstrations for programming human-robot collaboration (HRC) tasks. Recently, GhostAR presented a workflow of authoring HRC tasks by externalizing the human demonstration and using that as a time-space reference to program the robot collaborators. Further, Porfirio et al. [50] applied the method of human demonstration for human-robot social interactions.

To summarize, an embodied demonstration empowers rapid creation of complex and dynamic content through intuitive and straightforward bodily interactions. It is, therefore, suitable for machine task tutorial authoring especially in a fast-changing working environment. We envision the embodied demonstration to become the predominant method for creating machine task tutorials in future factory scenarios. While we apply this method for generating the tutor contents, we also emphasize the design space of the tutor presence in AR.

3 MACHINE TASK TUTORING

3.1 Machine Task: Local, Spatial, and Body-coordinated

This paper presents a study of AR presence for ***machine tasks*** tutoring system design. We define a machine task as a sequence of steps involving machine operations and spatial navigation, particularly for applications in production. A machine task is commonplace in workshop and factory environments, for the purposes of parts manufacturing, assembly, and equipment maintenance, repair, and overhaul. A ***step*** is the unit of a machine task sequence, which represents a meaningful inseparable action of the human-machine interaction. The steps in a machine task are usually a mixture of various types of interactions. In this paper, we focus on transferring knowledge regarding human actions. Therefore we elect to categorize the machine task steps by the level of movement required for the human-machine interaction. Based on our observation and engineering knowledge, as well as reviews from prior literature [58, 32], we classify the steps into the following three categories:

- A ***local*** step is a one-hand human-machine interaction in the user's current location and perspective. The user does not need body-scale spatial movements before interacting with the machine, nor does he/she need compound body-hands-eyes coordination for the action. Example *local step*s are simple actions with machine interfaces, such as with buttons, sliders, handles, knobs, and levers.
- A ***spatial*** step requires the user to perform noticeable spatial navigation before the machine interaction. The key challenge of this type of action is locating the target interface. Example *spatial step*s are tool change tasks during the machining operation that require the user to navigate to the designated area and find the right tool; or interactions with the machine interfaces that are away from the user's current location.
- A ***body-coordinated*** step is usually a two-hand action that requires the user to coordinate his/her body, hands, and eyes to complete the task. Example *body-coordinated step*s are the actions that operate two machine interfaces with two hands, respectively, in a synchronized or cooperative manner.

Figure 2 illustrates an example machine task using a band saw machine to manufacture a part. The user first needs to configure the machining parameters through a button and a knob, which are *local step*s (Figure 2-(1)). The user also needs to adjust the cutting angle and cutting-saw height using both hands in a coordinated manner, which is a *body-coordinated step* (Figure 2-(2)). Before starting the machine, he needs to

Figure 2. An example real-life machine task scenario involving *local* (1), *body-coordinated* (2), and *spatial* (3) interactions.

choose a base material meeting his production requirement from the material storage station, which is a *spatial step* (Figure 2-(3)). Note that in this study we focus on human-machine operations performed by the hands only, machine operations involving the feet are outside of the scope of this study.

3.2 Tutor Design from Embodied Authoring

When an apprentice is trying to learn a new machine task in a factory, the most effective way is to observe and follow the demonstration of an experienced master. We take the master-apprentice paradigm as an inspiration for our tutoring system design. The machine task tutorials in this paper are created from recording the physical demonstration of an expert (Figure 1-(1)), and it is displayed to users with the different visual presentations of the tutor (Figure 1-(2,3)). This paper focuses on exploring the tutor's *visual* representation only and does not include input to other senses, such as audio and tactile. We explore a design space of tutor presence in AR which involves spatial recording of the embodied authoring from an expert (i.e., the expert creates the tutorial by demonstrating the procedure).

Video. This tutor option mainly serves as a benchmark, since video is a popular tutoring media. To adapt video to fit our design space of embodied AR authoring, this option uses a video recording of the expert's first-person view while they demonstrate the task. The video recording is displayed to the user in a picture-in-picture style (Figure 1-(3)-a) at a fixed location and orientation in their visual field. Similar approaches have been used in prior work [23, 13].

Non-avatar-AR. This tutor option is similar to the existing AR instructions found in the machine-related tutoring systems discussed in our review of related work. It utilizes animated superimposed virtual models to represent the movement of the real part, aided with guiding symbols like arrows and text (Figure 1-(3)-b). A red circle on the ground indicates the spatial location of the tutor when they were recording. This tutor option represents the baseline of the existing AR instructions. A more detailed demonstration list for various machine interfaces can be found in Figure 3.

Half-body+AR. This tutor option displays an additional half-body avatar on top of the *non-Avatar-AR* option. The half-body avatar only has a visualization of the upper body and two armless hands, with the red circle indicating the ground position (Figure 1-(3)-c). Since all of the human-machine interactions in this paper are hands-only, we expect the virtual hands of this avatar to be sufficient for expressing the interaction. The head model indicates where to look and pay attention, while the upper-body plus the ground circle represent the spatial location of the tutor. This style of the avatar visualization

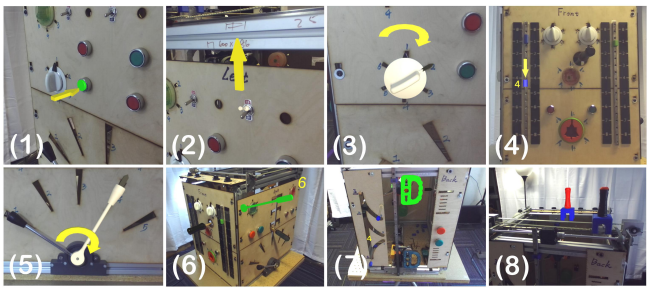

Figure 3. Example AR instructions for various machine interfaces: (1) button, (2) switch, (3) knob, (4) slider, (5) lever, (6) side-shift, (7) back-shift and (8) 2-DOF curve handle.

focuses on simplicity and is similar to the approaches used in prior research and commercial products [1, 60].

Full-body+AR. This tutor option displays an additional full-body avatar on top of the *non-avatar-AR* option. The avatar has a complete humanoid body structure, including arms and legs (Figure 1-(3)-d). Even though our tasks do not involve feet interactions, we choose the style of this avatar visualization due to its higher similarity to a real human tutor. The full-body avatar has already been widely adopted by prior work in various applications, such as ballet [59] and tennis training [45], Tai-Chi practice [27], MR remote collaboration [48], and telepresence meeting [46]. In our case, we are particularly interested in finding out whether and in what way the added avatar visualization would improve the user's understanding of the tutor's bodily movement.

Both the avatar tutor options include the AR instruction of the *non-avatar-AR*. While we agree on the necessity for intuitive and accurate instructions, our interests lie in understanding the effect of the added avatar visualization in the machine task tutoring scenarios. The four proposed tutor options represent the current mainstream AR-avatar related tutorial media. We design them to present the same instruction accurately while gradually incrementing their levels of guidance visualization. By studying the users' reactions under these four conditions, we aim to scale the weight of avatars in the AR tutoring systems and reveal the potential strengths and limitations of using them. Further, based on the study results, we seek the balance points in the level of visualization details for practical AR tutoring scenarios.

3.3 Implementation

Our see-through AR system is developed by attaching a stereo camera (ZED Dual 4MP Camera with a 2560 $\times$ 720 resolution at 60 fps and a field of view (FOV) of 90° (H) $\times$ 60° (V) $\times$ 110° (D) [57]) in front of a VR headset (Oculus Rift [1]), which is connected to a PC (Intel Core i7-9700K 3.6GHz CPU, 48GB RAM, NVIDIA GTX 1080). The positional tracking is enabled by four external sensors (Oculus IR-LED cameras), covering an effective area of 3 $\times$ 3m. To represent our *half-body* and *full-body* avatar, we choose a robotic humanoid avatar created by Noitom [43] due to its unbiased sexuality. We also adopt the hands model from Oculus Avatar SDK due to an expressive gesture visualization. Our system is developed using Unity3D (2018.2.16f1) [62] for both tutorial authoring and playback. The full-body avatar is estimated from the three-

point tracking (head and two hands) via inverse kinematics powered by a Unity3D plugin (FinalIK [19]).

4 EXPLORATORY USER STUDY

4.1 Study Setup: the Mockup Machine

In order to conduct our study, we first need to create a study scene to simulate machine tasks, that is capable of *local*, *spatial*, and *body-coordinated* human-machine interactions. We therefore created a mockup machine as the testbed for our study. The design of the mockup machine is guided by the following considerations: 1) The mockup machine should mimic real-life machine operation with realistic physical interfaces. 2) The size of the machine should be large enough to facilitate spatial navigation and bodily movement. 3) The machine should be designed with enough complexity to support the test sequence designed for the machine tasks. 4) Each interface on the mockup machine should provide multiple interaction possibilities in order to test and measure the user's performances.

Figure 4-Top illustrates the detailed design of our mockup machine ($0.7\times0.7\times0.7$ m), which is placed in the center of the study area (Figure 4-(f)), on top of a table (height = 0.78 m). The mockup machine can support *local* interactions via the following five interfaces: button, switch, knob, slider, and lever. It can support *spatial* interactions by asking the users to operate an interface on another side of the machine, which requires the users to first navigate spatially then locate the target interface before the interaction. We also designed a *spatial* 'key' interaction, simulating real-life tool change and assembly operation. In this interaction, users first need to go to the *key station* (Figure 4-(e)) and find the correct key, then walk back and insert it into a designated keyhole. As for *body-coordinated* interactions, we present a list of example interactions in Figure 4-(a-d). The first type of *body-coordinated* interaction supported by the mockup machine is operating two interfaces (slider-slider, slider-lever, lever-lever) with two hands respectively, in a synchronized manner (Figure 4-(a)). We've also specially designed three *body-coordinated* interfaces, including two 'shift' interfaces (Figure 4-(b,c)) that require user's both hands to operate in a cooperative manner; and a 'curve' interface requiring user's body-hands-eye coordination while operating the 2-DOF handle to repeat the trajectory in the tutorial (Figure 4-(d)).

4.2 Study Design

During the study, each user was asked to follow the tutor and complete four sessions of machine-operating task sequences. For each different session, the user followed a different tutor option to complete a different task sequence.

Sequence design. Each machine task sequence in the study consists of 36 steps, that are roughly evenly-distributed into three interaction categories: 1) *local* (10 steps including 2*button, 2*switch, 2*knob, 2*slider, and 2*lever), 2) *spatial* (14 steps including 2*button, 2*switch, 2*knob, 2*slider, 2*lever, and 4*'key'; these steps require large spatial navigation before interacting with the target interface), and 3) *body-coordinated* (12 steps including 2*slider-slider, 2*slider-lever, 2*lever-lever, 2*'side-shift', 2*'back-shift', and 2*'top-curve'). The

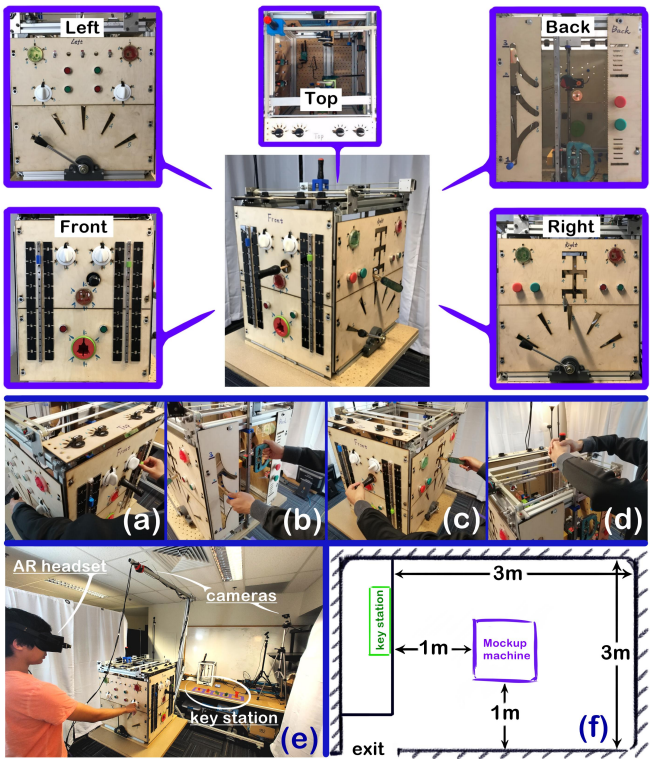

Figure 4. Top: The mockup machine detail design. Middle: example Body-coordinated machine interaction, including (a) two-interface synchronized operation, (b) back-shift, (c) side-shift, (d) top curve. Bottom: (e,f) study area setup layout.

four sequences are designed with the same step composition and execution order, to ensure the same task difficulty. To avoid memorization from previous sequences, the corresponding steps across different sequences have different detailed interactions. For example, step-1 on sequence-3 asks the user to twist the *right knob* to *position-3*, while the same step on sequence-4 asks the user to twist the *left knob* to *position-4* instead.

Tutorial length normalization. It's likely that the duration of a tutorial demonstration will affect users' task completion time. Since we created the tutorial for each of the four machine task sequences separately, the duration of corresponding steps across the different tasks are different. To enable a direct comparison of task completion time for corresponding steps across tasks, we scaled each step's duration to the average duration across the four corresponding steps, by slowing down or speeding up their playback. This procedure was performed for each set of four corresponding steps across the 36 steps in each sequence.

Data counterbalancing. To mitigate learning effects, the order in which participants used the different tutor options was counterbalanced across participants, such that each tutor option was tested on each ordinal position (first, second, third, fourth) with equal frequency. This was achieved by shuffling tutor options evenly with respect to the session order, resulting in a pre-arranged rotation list of 4 (sessions) * 4 (tutor options) = 16 participants. In total, we invited 16*2 = 32 participants for a balanced data acquisition.

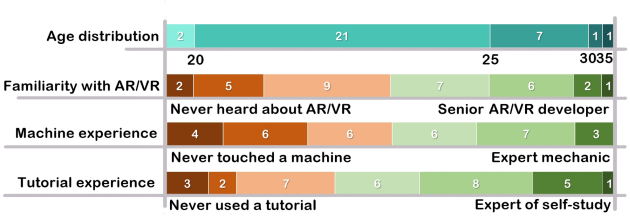

Figure 5. Demography of 32 participants.

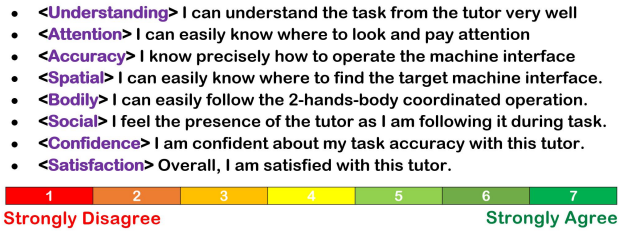

Figure 6. User experience survey questionnaire.

4.3 Participants

We recruited 32 users from our university via emails, posters, and networks (22 male and 10 female students between the ages of 18 and 35, $M = 23.8, SD = 3.64$). Each user was compensated $10. We did not particularly seek participants with AR/VR experience or machine operation skills for unbiased potential insights. We measured their familiarity with AR/VR on a 7-point Likert scale, with 1 being a total non-experienced user and 7 being an expert developer, yielding a result of $M = 3.63, SD = 1.43$. We also surveyed their general experience with hands-on interactions with machine-like objects ($M = 3.47, SD = 1.54$). Further, we asked users to rate their familiarity of self-teaching using any forms of tutorials ($M = 4.03, SD = 1.57$). An illustration of the user's demography survey results can be found in Figure 5.

4.4 Procedure

After completing the demographic survey, each user received a 5 minute introduction about the study background and a brief demonstration of how to interact with each interface on the mockup machine. The users then proceeded to the four sessions one by one, each session took the user about 10 minute to interact with the mockup machine and 5 minute afterward to fill out a user experience survey questionnaire. During each session, users were asked to wear the AR HMD and follow the machine task tutorial step by step. Users were asked to perform the machine operations at the comfortable speed of their choices, with no need to hurry or drag. A researcher monitored the entire process through the users' first-person AR view. If the researcher observed the user had completed the current step, he would switch to the next step and notify the user verbally. After completing the four sessions, users filled out a preference survey comparing the four tutor options, then finished up the study with a conversational interview.

4.5 Data Collection

Each user's study result contains three types of data: (1) tutorial following performance, (2) 7-point Likert subjective rating and user preference survey, and (3) conversational feedback.

Video Analysis. We recorded the entire study process using three cameras. The main source of objective data came from the video record of users' first-person AR view during the human-machine operation. We segmented this video into steps and manually coded the completion time and correctness of each step. Here we consider a step as completed if the user finished interacting with the interface and retrieved his/her hand. Also, we regard a step as completed correctly if the user interacted with the correct target interface and performed the correct positional manipulation (for slider, knob, lever, etc.), according to the corresponding tutor's demonstration. This yielded a total of 32(users)*4(sessions)*36(steps) = 4608 steps of objective analysis data across the entire study. We also had a top-view camera capturing the trajectory of the 'curve' interaction for accuracy analysis and a third-view camera recording from the top corner of the study scene for additional references.

Questionnaire. After each session, users rated their experience and subjective feelings for this session's tutor option using a 7-point Likert survey. The design of the survey question was derived from the standard user experience surveys, including *Single Ease Question (SEQ)* [53], *Subjective Mental Effort Question (SMEQ)* [54], *System Usability Scale (SUS)* [5], and *Networked Mind Measure of Social Presence (NMMSP)* [7], with added machine task elements and fine tuned specifically to our application scenario. The detailed questions are shown in Figure 6.

Interview. We audio-recorded all the subjective comments and suggestions from the users for post-study analysis and summary. During the study, we encouraged the users to 'Think Out-loud' to capture any on-the-fly insights as they were following the machine-operating tutorial. After the four sessions, we interviewed the users by asking their preference comparing all the tutor options for the machine task overall, as well as specifically for *local*, *spatial*, and *body-coordinated* interactions. The subjective feedback is later used in the paper to explain the study results and inspire for future design insights.

5 RESULTS

In this section, we present the results of this study. We first show the users' objective performances and subjective ratings, as well as tutor preferences. Then we provide a summary and explanatory analyses for the results using interview feedback and our observation.

5.1 Objective Performance

We first demonstrate the overall user performance by comparing four different tutor options. Then we present detailed user performances regarding each interaction category: *local*, *spatial*, and *body-coordinated*. We measure the completion time and accuracy, which reveals how efficiently and accurately the users understand the tutorials. Since the tutorial for each step has a different duration, we normalize the completion time of each step as: actual step completion time divided by the duration of the step demonstration in the tutorial. The

tutorial duration for a machine task sequence (36 steps) is: 6 minutes 15 seconds, with each step's length ranges between 4.9 to 19.3 seconds, while the average completion time of a sequence is: 7 minutes 21 seconds. The accuracy of a category of steps is calculated as: the number of correct steps divided by the total step number. To characterize the accuracy of the 2D 'curve' operation, we calculate the Modified Hausdorff Distance (MHD) [16] between the trajectory performed by the user and the one in the corresponding tutorial, with a smaller distance indicating more similarity and higher accuracy. The normal distribution assumption is violated by our dataset as indicated by Shapiro-Wilk normality test ($p < 0.005$). Hence to examine the statistical significance across the four tutorial options, we conduct a Friedman test with a Wilcoxon signed-rank, rather than the repeated ANOVA measures. All results are presented in Figure 7.

Overall performance. The average normalized completion time shows that the users spend the longest amount of time following the *video* tutorials ($M = 1.58, SD = 0.70$) and is significantly slower than *non-avatar-AR* tutor option ($M = 1.16, SD = 0.57$) ($Z = -18.416, p < 0.0005$). Among all the AR options, the *half-body+AR* tutorials ($M = 1.14, SD = 0.55$) shows marginally shorter completion time than *non-avatar-AR* ones, with no significant edge ($Z = -0.854, p = 0.393$). Meanwhile, users with *full-body+AR* tutorials ($M = 1.15, SD = 0.47$) perform slightly slower than the ones with *half-body+AR* ($Z = -2.527, p < 0.05$). The accuracy result reveals the same trend as the completion time. The *video* tutorials has the lowest accuracy ($M = 85.4\%, SD = 6.28\%$) while the accuracy of *non-avatar-AR* ($M = 95.6\%, SD = 3.82\%$), *half-body+AR* ($M = 96.3\%, SD = 2.82\%$) and *full-body+AR* ($M = 95.8\%, SD = 2.87\%$) options are approximately equally high (pairwise $p > 0.05$).

***Local* steps performance**. Similar to the overall performance, the *video* option still has the poorest performance in terms of task completion time ($M = 1.09, SD = 0.21$) and accuracy ($M = 92.5\%, SD = 3.36\%$). Interestingly, users with *non-avatar-AR* tutor option ($M = 0.80, SD = 0.24$) are significantly faster than the ones with *half-body+AR* tutor ($M = 0.87, SD = 0.23$) and *full-body+AR* tutor ($M = 0.93, SD = 0.26$) ($Z = -4.487, p < 0.0005$ and $Z = -6.189, p < 0.0005$ respectively), which implies that the existence of an avatar may have negative influence on the user's perception for *local* task understanding. In terms of the accuracy, no significant difference was found among *non-avatar-AR* ($M = 99.1\%, SD = 1.48\%$), *half-body+AR* ($M = 97.5\%, SD = 2.54\%$), and *full-body+AR* ($M = 98.4\%, SD = 1.84\%$) (pairwise $p > 0.05$).

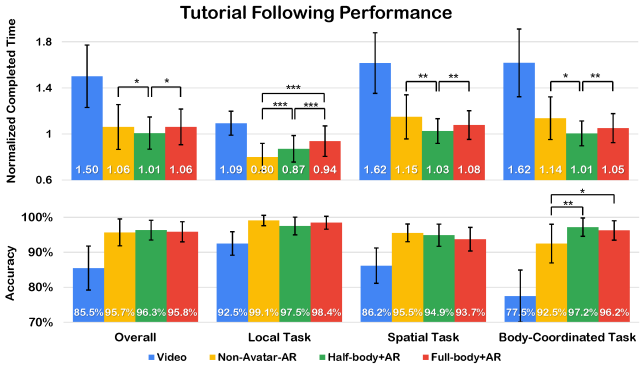

Figure 7. Tutorial following performance. (*=p<.0005, **=p<.005, *=p<.05. If not specified, *** between the video options and other three tutor options.) Error bars represent standard deviations.**

***Spatial* steps performance**. The *video* option takes the longest time to complete ($M = 1.62, SD = 0.52$) and receives the lowest accuracy ($M = 86.2\%, SD = 5.05\%$). While the *half-body+AR* tutorials achieves relatively shorter completion time ($M = 1.02, SD = 0.22$) than *non-avatar-AR* ($M = 1.15, SD = 0.38$) and *full-body+AR* ($M = 1.07, SD = 0.25$) ($Z = -2.19, p < 0.028$ and $Z = -2.750, p < 0.006$ respectively). On the other hand, the accuracy of *non-avatar-AR* ($M = 95.5\%, SD = 2.53\%$), *half-body+AR* ($M = 94.9\%, SD = 3.17\%$), and *full-body+AR* ($M = 93.7\%, SD = 3.36\%$) are roughly the same (pairwise $p > 0.05$).

***Body-coordinated* steps performance**. The *video* tutorials received the worst performance in both completion time ($M = 1.62, SD = 0.58$) and accuracy ($M = 77.5\%, SD = 7.4\%$). Users with *half-body+AR* tutor option (normalized completion time: $M = 1.00, SD = 0.22$, accuracy: $M = 97.2\%, SD = 2.61\%$) are able to perform significantly faster ($Z = -2.19, p < 0.05$) with less mistakes ($p < 0.05$) than the ones with *non-avatar-AR* tutorials (normalized completion time: $M = 1.13, SD = 0.37$, accuracy: $M = 92.4\%, SD = 5.54\%$), which indicates the strengths of the avatar in demonstrating bodily movement. Between the two avatar options, the *full-body+AR* (normalized completion time: $M = 1.05, SD = 0.25$, accuracy: $M = 96.2\%, SD = 2.77\%$) has longer completion time ($Z = -2.75, p = 0.005$) and roughly the same accuracy ($p > 0.05$) compared with *half-body+AR* tutor option. For the 2D 'curve' operation, the *half-body+AR* tutor achieves the shortest average MHD ($M = 42.5\ cm, SD = 13.7\ cm$), followed by *non-avatar-AR* ($M = 46.1\ cm, SD = 21.0\ cm$) and *full-body+AR* ($M = 46.6\ cm, SD = 20.0\ cm$), while the *video* tutor option achieves the longest average MHD ($M = 50.7cm, SD = 18.25cm$). Yet the Friedman test ($\chi^2(3) = 5.81, p = 0.121$) does not reveal any significant difference among the four options.

5.2 Subjective Rating and User Preference

Figure 8 shows the user experience subjective ratings with the 7-point Likert questionnaire. To reveal the differences among the tutor options, we conduct a Friedman test followed by a Wilcoxon signed-rank test on each of the eight questions individually. We first look into the effectiveness of AR in the tutorial systems by comparing *video* and *non-avatar-AR*. The result shows that the latter option achieves significantly higher ratings ($p < 0.0005$) in 'Understanding', 'Accuracy', 'Confidence' and 'Satisfaction', while no significant difference is found in 'Attention' ($p = 0.22$), 'Spatial' ($p = 0.124$), 'Bodily' ($p = 0.167$) and 'Social' ($p = 0.355$). Secondly, we examine whether the existence of an avatar affects the user experience. The result reveals that in all eight ratings, the *non-avatar-AR* option has significant lower scores ($p < 0.05$) than either the *half-body+AR* or the *full-body+AR*. Thus, we believe the overall machine task user experience is improved

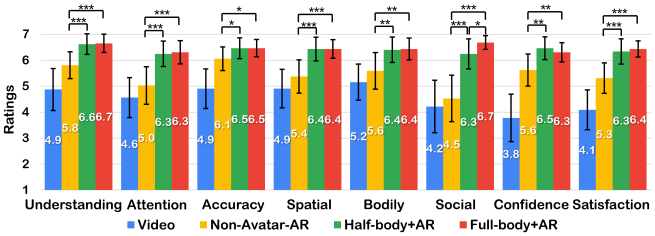

Figure 8. User experience ratings. (*=p<.0005, **=p<.005, *=p<.05. If not specified, *** between the video options and other three tutor options.) Error bars represent standard deviations.**

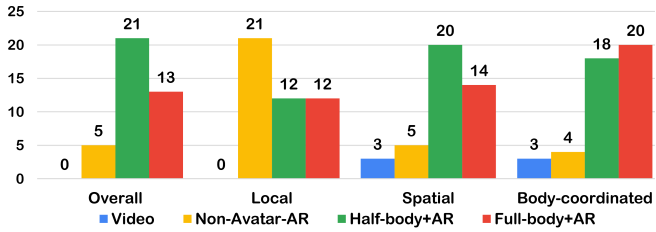

Figure 9. User preference result.

by the presence of an avatar. Finally, we inspect how the visual guidance level of avatar influences the user experience by comparing the ratings between *half-body+AR* and *full-body+AR*. We find no significant difference between the two tutor options except the 'Social' rating where *full-body+AR* is slightly higher ($p = 0.05$) than *half-body+AR*.

Figure 9 illustrates the user preference survey results for the overall machine task tutoring experience, regarding the *local*, *spatial*, and *body-coordinated* interactions, respectively. For each type of interaction, users are allowed to choose one or more tutor options as their favorite. Overall, the *half-body+AR* is most preferred tutor option (21 out of 39), followed by the *full-body+AR* (13 out of 39) and the *non-avatar-AR* (5 out of 39), while no users choose the *video* as their favorite tutor option. In terms of the *local* interactions, the *non-avatar-AR* option is the most favored (21 out of 45). The *half-body+AR* and the *full-body+AR* are tied in the second place (12 out of 45). Again, no users choose the *video*. In terms of the *spatial* interactions, the *half-body+AR* tutor option comes to the first place (20 out of 42). The *full-body+AR* option takes the second place with 14 out of 42 users, while *non-avatar-AR* (5 out of 42) and *video* (3 out of 42) are less preferred. As for the *Body-coordinate* interactions, *full-body+AR* is the most popular choice (20 out of 45), that is followed closely by *half-body+AR* (18 out of 45). Only a few users choose the *non-avatar-AR* (4 out of 45) and the *video* (3 out of 45) tutor option as their favorite.

5.3 Result Summary and Analysis

We now summarize the main results and present explanatory analysis using our observation during the study as well as findings that come out from the interview.

5.3.1 Overall favorite: half-body vs. full-body

The ratings for the two proposed avatar tutor options are found to be similar across all categories, and are significantly better than the *non-avatar-AR* and the *video* options (Figure 8). Interestingly, when the users' are asked to pick their favorite tutor option overall, the *half-body* has a clear preference edge over the *full-body* (21 vs 13). From the post-study interview, we find that many users believe these two tutor options are functionally equal, while the *half-body* has less occlusion to the users' views. "*I think the half-body is the best because it can show me where to go and what to do without blocking too much of my sight (P7).*" The increased visual access to the physical machine in *half-body* as compared to *full-body* may also have resulted in lower mental effort ("*The full-body avatar tutor shows too many things, and sometimes is too exhausting for me (P8)*"), and less attention distraction ("*A full-body human is not necessary, its arms and legs distract my attention from the machine, half-body is cleaner and less distracting (P16)*"). This is also reflected in the objective performance result (Figure 7) where the *half-body* achieves similar accuracy with less time, compared with the *full-body*. The above discussion also explains the preference result for the *spatial* interactions, where the *half-body* is enough for instructing spatial navigation and target finding, with a cleaner observing view.

As observed from the user preference result, the additional body features become helpful in the *body-coordinated* interactions. The users feel that the added limb representations, especially the arms, do provide a better understanding of the two-hand coordinated tasks. "*For the bodily tasks, I prefer full-body, because full-body gives me more spatial and embodied evidence. Just a hand is not enough sometimes. I feel like needing the extra arm information (P15).*" Another preference of the *full-body* over the *half-body* is on social presence, which is also reflected by the 'Social' subjective rating results. According to the feedback from the interview, the *full-body* is better than the *half-body* at representing a human tutor, which makes it a more friendly, believable, and reliable option. " *The full-body feels more like a human, like a real tutor and more friendly. In comparison, the half-body is obviously a robotic indicator (P26).*"

5.3.2 Local favorite: non-avatar-AR

Despite the lack of spatial and bodily presentations, the *non-avatar-AR* is selected as the favorite tutor option for the *local* interactions. This is because the *local* interactions does not require substantial spatial and body movements. The attention of the user does not need to be directed effectively to locate the target interface, nor does a particular body gesture play an important role in terms of interaction execution. Therefore, the presence of a human avatar in *local* interactions does not provide extra benefits in most cases. "*I don't think avatar is useful for local tasks because AR instruction is enough, and I usually cannot see the avatar anyways because I am standing inside the avatar (P5).*" Several participants report that the avatar encumbers and slows down their actions, which is consistent with our finding that the *non-avatar-AR* is fastest for *local* tasks with equal accuracy (Figure 7).

5.3.3 Least preferred: video

It is clear that the *video* is the least popular tutor option among the four. According to our observation, the main problem for *video* tutoring is caused by the two separate dimensions

of the tutoring and the application: users have to receive the instruction from the digital world, interpret it into his/her physical world, and then apply it to the corresponding machine interfaces. This translation gap causes many problems such as distracted attention, fractured spatial mapping, high mental effort from memorization, and a non-optimized observation perspective. "*My attention is changing from video content to reality all the time, and sometimes I need to think very hard to interpret what it means in the video (P1).*" However, the *video* still demonstrates some values from the user preference survey on *spatial* and *body-coordinated* categories. Some users have mentioned that the *video* option can occasionally be more expressive than the other options. "*To me, video is the best for spatial and embodied task, because you can best understand the body motion right away. The avatar is not obvious because I was standing inside the avatar, and I cannot notice the avatar (P2).*"

6 DISCUSSION

In this section we discuss the primary results of the study and contrast them with prior works. We also provide design recommendations and insights for future AR tutoring systems.

6.1 Benefits of Avatars for Tutoring

Our first research question focuses on understanding whether the proposed AR avatar presentations improve the machine task tutoring experience, and how they do so. Our findings indicate that the AR avatars receive significantly more positive feedback than the non-avatar and video tutor options, and provide several insights into why. We summarize these reasons below, and distill our findings into design recommendations for avatar-based tutoring systems.

Spatial Attention Allocation. When trying to follow a comprehensive machine task tutorial, one of the major challenges for a user is to know where to pay attention, especially during constant spatial movements that easily cause disorientation. Compared to the non-avatar tutor presence, the additional avatar provides more noticeable in-situ visual hints to guide the user's attention. "*Sometimes I cannot find the machine target until the human avatar moves over there and starts reaching out his hand. (P7)*" This result is aligned with prior works on mixed reality assistant, where a remote expert provides *live* guidance for a local learner via the presence of an AR avatar [60, 49, 48]. The avatars in these works are usually controlled by a remote human, thus are capable of communicating and responding to the user's action adaptively. However, when applying the avatars to *recorded* tutoring with no remote human involvement, we recommend that future system provide a feedback mechanism for user-responsive tutoring. For example, the *recorded* tutor should act only when the learner is paying attention to it [39]. The above finding also inspires us to design attention indicators in the future, that can explicitly guide users' attention and reduce mental effort.

Bodily Movement Expression. The digital tutor is capable of intuitively expressing the human body in the context of a physical interactive target. This enables the users to understand the movement accurately and anticipate the tutor's actions, especially for the tasks involving head-hand-body coordination. "*Seeing the human tutor move in the space allows me to predict where he is going and what he is going to do next, and it prepares me to get ready for the task in advance (P20).*" This advantage of avatars is consistent with prior research on body movement training, such as the YouMove system [3]. While prior works on body movement training [3, 42, 11] have primarily focused on physical tasks being performed by humans in isolation, we show that these advantages have benefits for tasks where spatial and temporal connections must be made between virtual avatars and physical objects (in our case, the machine being manipulated).

Higher Social Presence. Due to the human-like visual presence, user feedback suggests that following the avatar resembles the tutoring experience of following a human teacher. This improves the user's confidence, which leads to a higher efficiency in tutoring information transfer. "*The human avatar is easy to follow, as long as you do that, you feel confident, and nothing is going to be wrong, it gives me less mental pressure (P24).*" Mini-Me [48] has a similar finding that the avatar option in their study yields a higher aggregated social presence and awareness score for task transfer collaboration than the non-avatar options, resulting in the reduced mental effort and improved performance.

6.2 Adaptive Tutoring

In the second research question, we explore how to optimize the tutoring experience in a comprehensive machine task scenario involving multiple interaction categories. In this paper, we study four tutor options with gradually increased guidance visualization level, aiming to provide insights for the ideal design. Our results do not show any one presentation method to be clearly superior, but rather reveal a number of considerations that must be balanced to create a good avatar-based tutoring experience. In particular, we discuss three factors in the sections that follow: level of visual detail, tutor following paradigm, and playback progress. As some of these factors reflect individual preferences, we believe there is an opportunity for adaptive and personalized tutoring experiences that dynamically tailor the experience to individual users.

Level of Visual Detail. According to our results, users acknowledge the usefulness of avatars, but more visual details also cause confusion and occlusion of the physical world. Therefore the level of visual guidance details should be contextually adaptive to the interaction type and task difficulty. This also explains why the users prefer half-body avatar for the overall machine task and non-avatar for the *local* interactions. "*It should not display the whole action animation, only the key part should be played; otherwise, it is too distracting (P1).*" This finding aligns with a study conducted by Lindlbauer D, et al. [39] where they find that the dynamically adjusted AR contents lead to less distraction and higher performance. Further, the tutor's presence should also adapt to the learner's reliance on instructions. We have observed that some users were able to complete most steps fast and accurately by only following *non-avatar-AR* option, while some others needed *full-body+AR* instead.

Tutor Following Paradigm. Currently, the position of the AR tutor is connected to the physical interactive machine, and it is up to the learners to decide where to observe the tutor and how to follow it. We noticed that some users prefer to stand inside the tutor avatar and follow it in synchronization to achieve higher efficiency and accuracy. "*I like to stand inside of the avatar and follow its movement, makes me feel confident about my accuracy (P29).*" This paradigm has been acknowledged and adopted by an arm motion training system where the virtual guiding arms are superimposed in the user's egocentric AR view [26]. On the other hand, some users prefer to stand on the side of the avatar tutor because they consider it uncomfortable to collide into a virtual humanoid. "*I do not feel like standing inside the avatar, because it feels like a person and I don't want to crash into him (P6).*" This can be explained by a study conducted by Kim et al. [35] on the physical presence of the avatar. They find that the conflicts between humans and virtual avatars reduce the sense of co-presence and should be avoided if possible. The above findings demonstrate the importance of providing spatially aware instructional contents based on the user's physical location and observation perspective.

Playback Progress. In our study setup, the playback speed of each tutorial step is fixed and determined by the authored demonstration. Also, the progress of the user is manually monitored and manipulated by the researchers. If a learner misses critical information of the step, he/she has to wait for the step animation to play again, leading to low learning efficiency. Based on our observation and feedback, we believe the future systems should incorporate an adaptive tutorial playback speed based on users' innate capability and task difficulty. This finding is aligned with the study done by Rajinder et. al [56], where they study projected visualizations for hand movement guidance and find that dynamically adjusted guiding speed has the potential of improving training efficiency. Further, an adaptive playback helps the users to preview the tutor's intent, such as using slow-motion to forecast the avatar's actions. "*I need to know what the avatar is about to do and where to pay attention, sometimes the avatar makes a sudden turn, and it's very hard to notice (P23).*"

7 STUDY LIMITATIONS

The hardware and performance of the AR headset may have influenced participants' experience in several ways. Though we used state-of-the-art technology (VR headset with a front-attached stereo camera to achieve see-through AR with a high-resolution and full eye-sight field of view), several participants reported minor motion sickness, and the inability for the cameras to fully simulate stereo vision that caused some participants to bump into the machine while trying to manipulate it. As the headset was tethered to a computer, cords sometimes needed to be untangled, which may have slowed spatial and bodily movements as compared to operating the machine free from tethers. While we acknowledge that the above conditions may have impacted the user experience, they were consistent across the three tutor options we tested.

To conduct our study, we have created an interactive mockup machine capable of all three types of steps. Therefore the study result that we collected is largely based on the users' interaction performance on this mockup machine. Even though we designed the mockup machine based on the real-world machine interfaces and interactions, it is still a testbed. The mockup machine can only represent a portion of the real-world machine tasks, including the three interaction steps. We would like to acknowledge explicitly that the result of this study should be used mainly as a comparative reference among the four tutor options as an elicitation or informative study for future tutoring system design.

8 CONCLUSION

In this paper, we have presented an exploratory study of augmented reality presence for machine task tutoring system design. We created an AR-based embodied authoring system capable of creating tutorials with four types of tutor options: *video*, *non-avatar-AR*, *half-body+AR*, and *full-body+AR*. In order to conduct our study, we have designed and fabricated a mockup machine capable of supporting *local, spatial*, and *body-coordinated* human-machine interactions. We invited 32 users, each for a 4-session study experiencing all four tutor options for comparative feedback. From the quantitative and qualitative results of the study, we have discussed and summarized the design recommendations for future tutoring systems. These design insights form an important stepping stone to help the future researchers create a comprehensive and intelligent machine task tutoring system, that will enable fluid machine task skill transfer and empower an efficient, flexible, and productive workforce.

ACKNOWLEDGEMENT

We wish to give a special thanks to the reviewers for their invaluable feedback. This work is partially supported by the NSF under grants FW-HTF 1839971 and OIA 1937036. We also acknowledge the Feddersen Chair Funds. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

[1] 2019. Oculus. (2019). `https://www.oculus.com/`.

[2] Heni Ben Amor, Gerhard Neumann, Sanket Kamthe, Oliver Kroemer, and Jan Peters. 2014. Interaction primitives for human-robot cooperation tasks. In *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2831–2837.

[3] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 311–320.

[4] armedia. 2019. I-Mechanic, the AR App that turns yourself into a Mechanic. (2019). Retrieved September 1, 2019 from `http://www.armedia.it/i-mechanic`.

[5] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[6] Connelly Barnes, David E Jacobs, Jason Sanders, Dan B Goldman, Szymon Rusinkiewicz, Adam Finkelstein, and Maneesh Agrawala. 2008. Video puppetry: a performative interface for cutout animation. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 124.

[7] Frank Biocca, C. Harms, and Jennifer Gregg. 2001. The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity. *4th annual International Workshop on Presence, Philadelphia* (01 2001).

[8] Jean-Rémy Chardonnet, Guillaume Fromentin, and José Outeiro. 2017. Augmented reality as an aid for the use of machine tools. *Res. & Sci. Today* 13 (2017), 25.

[9] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: automatic generation of step-by-step mixed media tutorials. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 93–102.

[10] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmot Li, and Bjoern Hartmann. 2013. Democut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 141–150.

[11] Pei-Yu Peggy Chi, Daniel Vogel, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2016. Authoring illustrations of human movements by iterative physical demonstration. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 809–820.

[12] Philo Tan Chua, Rebecca Crivella, Bo Daly, Ning Hu, Russ Schaaf, David Ventura, Todd Camill, Jessica Hodgins, and Randy Pausch. 2003. Training for physical tasks in virtual environments: Tai Chi. In *IEEE Virtual Reality, 2003. Proceedings.* IEEE, 87–94.

[13] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. 2016. You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding* 149 (2016), 98–112.

[14] Francesca De Crescenzio, Massimiliano Fantini, Franco Persiani, Luigi Di Stefano, Pietro Azzari, and Samuele Salti. 2010. Augmented reality for aircraft maintenance training and operations support. *IEEE Computer Graphics and Applications* 31, 1 (2010), 96–101.

[15] Gino Dini and Michela Dalle Mura. 2015. Application of augmented reality techniques in through-life engineering services. *Procedia Cirp* 38 (2015), 14–23.

[16] M-P Dubuisson and Anil K Jain. 1994. A modified Hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition*, Vol. 1. IEEE, 566–568.

[17] Daniel Eckhoff, Christian Sandor, Christian Lins, Ulrich Eck, Denis Kalkofen, and Andreas Hein. 2018. TutAR: augmented reality tutorials for hands-only procedures. In *Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. ACM, 8.

[18] Allen J Fairchild, Simon P Campion, Arturo S García, Robin Wolff, Terrence Fernando, and David J Roberts. 2016. A mixed reality telepresence system for collaborative space operation. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (2016), 814–827.

[19] FinalIK. 2019. FinalIK. (2019). Retrieved September 1, 2019 from `https://assetstore.unity.com/packages/tools/animation/final-ik-14290`.

[20] Markus Funk. 2016. Augmented reality at the workplace: a context-aware assistive system using in-situ projection. (2016).

[21] Dominic Gorecky, Mohamed Khamis, and Katharina Mura. 2017. Introduction and establishment of virtual training in the factory of the future. *International Journal of Computer Integrated Manufacturing* 30, 1 (2017), 182–190.

[22] Dominic Gorecky, Mathias Schmitt, Matthias Loskyll, and Detlef Zühlke. 2014. Human-machine-interaction in the industry 4.0 era. In *2014 12th IEEE international conference on industrial informatics (INDIN)*. Ieee, 289–294.

[23] Michihiko Goto, Yuko Uematsu, Hideo Saito, Shuji Senda, and Akihiko Iketani. 2010. Task support system by displaying instructional video onto AR workspace. In *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 83–90.

[24] Ankit Gupta, Maneesh Agrawala, Brian Curless, and Michael Cohen. 2014. Motionmontage: A system to annotate and combine motion takes for 3d animations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2017–2026.

[25] Ankit Gupta, Dieter Fox, Brian Curless, and Michael Cohen. 2012. DuploTrack: a real-time system for authoring and guiding duplo block assembly. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 389–402.

[26] Ping-Hsuan Han, Kuan-Wen Chen, Chen-Hsin Hsieh, Yu-Jie Huang, and Yi-Ping Hung. 2016. Ar-arm: Augmented visualization for guiding arm movement in the first-person perspective. In *Proceedings of the 7th Augmented Human International Conference 2016*. ACM, 31.

[27] Ping-Hsuan Han, Yang-Sheng Chen, Yilun Zhong, Han-Lei Wang, and Yi-Ping Hung. 2017. My Tai-Chi coaches: an augmented-learning tool for practicing Tai-Chi Chuan. In *Proceedings of the 8th Augmented Human International Conference*. ACM, 25.

[28] Ping-Hsuan Han, Jia-Wei Lin, Chen-Hsin Hsieh, Jhih-Hong Hsu, and Yi-Ping Hung. 2018. tARget: limbs movement guidance for learning physical activities with a video see-through head-mounted display. In *ACM SIGGRAPH 2018 Posters*. ACM, 26.

[29] Robert Held, Ankit Gupta, Brian Curless, and Maneesh Agrawala. 2012. 3D puppetry: a kinect-based interface for 3D animation.. In *UIST*. Citeseer, 423–434.

[30] Steven J Henderson and Steven K Feiner. 2011. Augmented reality in the psychomotor phase of a procedural task. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 191–200.

[31] Thuong N Hoang, Martin Reinoso, Frank Vetere, and Egemen Tanin. 2016. Onebody: remote posture guidance system using first person view in virtual environment. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, 25.

[32] Jean-Michel Hoc. 2001. Towards a cognitive approach to human–machine cooperation in dynamic situations. *International journal of human-computer studies* 54, 4 (2001), 509–540.

[33] Geun-Sik Jo, Kyeong-Jin Oh, Inay Ha, Kee-Sung Lee, Myung-Duk Hong, Ulrich Neumann, and Suya You. 2014. A unified framework for augmented reality and knowledge-based systems in maintaining aircraft. In *Twenty-Sixth IAAI Conference*.

[34] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 4017–4026.

[35] Kangsoo Kim, Gerd Bruder, and Greg Welch. 2017. Exploring the effects of observed physicality conflicts on real-virtual human interaction in augmented reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. ACM, 31.

[36] Seungwon Kim, Gun Lee, Weidong Huang, Hayun Kim, Woontack Woo, and Mark Billinghurst. 2019. Evaluating the Combination of Visual Communication Cues for HMD-based Mixed Reality Remote Collaboration. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 173.

[37] Yongkwan Kim and Seok-Hyung Bae. 2016. SketchingWithHands: 3D sketching handheld products with first-person hand posture. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 797–808.

[38] Bokyung Lee, Minjoo Cho, Joonhee Min, and Daniel Saakes. 2016. Posing and acting as input for personalizing furniture. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, 44.

[39] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 147–160.

[40] Matthias Loskyll, Ines Heck, Jochen Schlick, and Michael Schwarz. 2012. Context-based orchestration for control of resource-efficient manufacturing processes. *Future Internet* 4, 3 (2012), 737–761.

[41] Alejandro Monroy Reyes, Osslan Osiris Vergara Villegas, Erasmo Miranda Bojórquez, Vianey Guadalupe Cruz Sánchez, and Manuel Nandayapa. 2016. A mobile augmented reality system to support machinery operations in scholar environments. *Computer Applications in Engineering Education* 24, 6 (2016), 967–981.

[42] Christian Murlowski, Florian Daiber, Felix Kosmalla, and Antonio Krüger. 2019. Slackliner 2.0: Real-time Training Assistance through Life-size Feedback. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, INT012.

[43] Noitom. 2019. Perception Neuron by Noitom. (2019). Retrieved September 1, 2019 from `https://neuronmocap.com`.

[44] SK Ong and ZB Wang. 2011. Augmented assembly technologies based on 3D bare-hand interaction. *CIRP annals* 60, 1 (2011), 1–4.

[45] Masaki Oshita, Takumi Inao, Tomohiko Mukai, and Shigeru Kuriyama. 2018. Self-training system for tennis shots with motion feature assessment and visualization. In *2018 International Conference on Cyberworlds (CW)*. IEEE, 82–89.

[46] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. ACM, 1716–1725.

[47] Amaury Peniche, Christian Diaz, Helmuth Trefftz, and Gabriel Paramo. 2012. Combining virtual and augmented reality to improve the mechanical assembly training process in manufacturing. In *American Conference on applied mathematics*. 292–297.

[48] Thammathip Piumsomboon, Gun A Lee, Jonathon D Hart, Barrett Ens, Robert W Lindeman, Bruce H Thomas, and Mark Billinghurst. 2018. Mini-me: an adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. ACM, 46.

[49] Thammathip Piumsomboon, Gun A Lee, Andrew Irlitti, Barrett Ens, Bruce H Thomas, and Mark Billinghurst. 2019. On the Shoulder of the Giant: A Multi-Scale Mixed Reality Collaboration with 360 Video Sharing and Tangible Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 228.

[50] David Porfirio, Evan Fisher, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. 2019. Bodystorming Human-Robot Interactions. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*.

[51] D Preuveneers. 2015. The GhostHands UX: telementoring with hands-on augmented reality instruction. In *Workshop Proceedings of the 11th International Conference on Intelligent Environments*, Vol. 19. IOS Press, 236.

[52] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 622.

[53] Jeff Sauro. 2012. 10 THINGS TO KNOW ABOUT THE SINGLE EASE QUESTION (SEQ). (2012). Retrieved September 1, 2019 from `https://measuringu.com/seq10/`.

[54] Jeff Sauro and Joseph S Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1599–1608.

[55] Eldon Schoop, Michelle Nguyen, Daniel Lim, Valkyrie Savage, Sean Follmer, and Björn Hartmann. 2016. Drill Sergeant: Supporting physical construction projects through an ecosystem of augmented tools. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1607–1614.

[56] Rajinder Sodhi, Hrvoje Benko, and Andrew Wilson. 2012. LightGuide: projected visualizations for hand movement guidance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 179–188.

[57] Stereolabs. 2019. ZED Mini Stereo Camera - Stereolabs. (2019). Retrieved September 1, 2019 from `https://www.stereolabs.com/zed-mini/`.

[58] Lucy Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press.

[59] Guoyu Sun, Paisarn Muneesawang, M Kyan, Haiyan Li, Ling Zhong, Nan Dong, Bruce Elder, and Ling Guan. 2014. An advanced computational intelligence system for training of ballet dance in a cave virtual reality environment. In *2014 IEEE International Symposium on Multimedia*. IEEE, 159–166.

[60] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019a. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 161–174.

[61] Balasaravanan Thoravi Kumaravel, Cuong Nguyen, Stephen DiVerdi, and Björn Hartmann. 2019b. TutoriVR: A Video-Based Tutorial System for Design Applications in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 284.

[62] Unity. 2019. Unity Real-Time Development Platform. (2019). Retrieved September 1, 2019 from `https://unity.com/`.

[63] David Vogt, Simon Stepputtis, Steve Grehl, Bernhard Jung, and Heni Ben Amor. 2017. A system for learning continuous human-robot interactions from human-human demonstrations. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2882–2889.

[64] VRChat. 2019. Create And Play in Virtual Worlds. (2019). Retrieved September 1, 2019 from `https://www.vrchat.net/`.

[65] Sabine Webel, Uli Bockholt, Timo Engelke, Nirit Gavish, Manuel Olbrich, and Carsten Preusche. 2013. An augmented reality training platform for assembly and maintenance skills. *Robotics and Autonomous Systems* 61, 4 (2013), 398–403.

[66] Giles Westerfield. 2012. Intelligent augmented reality training for assembly and maintenance. (2012).

[67] Tomasz Wójcicki. 2014. Supporting the diagnostics and the maintenance of technical devices with augmented reality. *Diagnostyka* 15, 1 (2014), 43–47.

[68] Shuo Yan, Gangyi Ding, Zheng Guan, Ningxiao Sun, Hongsong Li, and Longfei Zhang. 2015. OutsideMe: Augmenting Dancer's External Self-Image by Using A Mixed Reality System. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 965–970.

[69] Yupeng Zhang, Teng Han, Zhimin Ren, Nobuyuki Umetani, Xin Tong, Yang Liu, Takaaki Shiratori, and Xiang Cao. 2013. BodyAvatar: creating freeform 3D avatars using first-person body gestures. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 387–396.

[70] J Zhu, Soh-Khim Ong, and Andrew YC Nee. 2015. A context-aware augmented reality assisted maintenance system. *International Journal of Computer Integrated Manufacturing* 28, 2 (2015), 213–225.

[71] Zhiwei Zhu, Vlad Branzoi, Michael Wolverton, Glen Murray, Nicholas Vitovitch, Louise Yarnall, Girish Acharya, Supun Samarasekera, and Rakesh Kumar. 2014. AR-mentor: Augmented reality based mentoring system. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 17–22.