Understanding Effects of Visual Feedback Delay in AR on Fine Motor Surgical Tasks

Talha Khan, Toby S. Zhu, Thomas Downes, Lucille Cheng, Nicolás M. Kass, Edward G. Andrews, Jacob T. Biehl

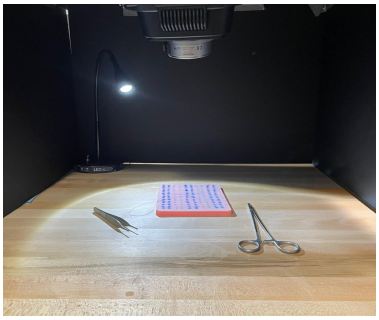
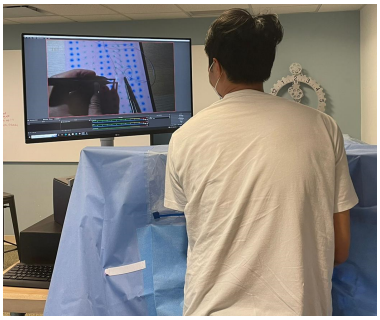
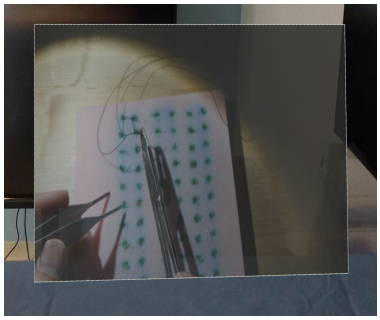

Fig. 1: **(Left)** Contents inside the box **(Middle)** Surgeon performing a task using the monitor **(Right)** Surgeon's view from HoloLens when performing a task

Abstract—Latency is a pervasive issue in various systems that can significantly impact motor performance and user perception. In medical settings, latency can hinder surgeons' ability to quickly correct movements, resulting in an experience that doesn't align with user expectations and standards of care. Despite numerous studies reporting on the negative effects of latency, there is still a gap in understanding how it impacts the use of augmented reality (AR) in medical settings. This study aims to address this gap by examining how latency impacts motor task performance and subjective perceptions, such as cognitive load, on two display types: a monitor display, traditionally used inside an operating room (OR), and a Microsoft HoloLens 2 display. Our findings indicate that both level of latency and display type impact motor performance, and higher latencies on the HoloLens result in relatively poor performance. However, cognitive load was found to be unrelated to display type or latency, but was dependent on the surgeon's training level. Surgeons did not compromise accuracy to gain more speed and were generally well aware of the latency in the system irrespective of their performance on task. Our study provides valuable insights into acceptable thresholds of latency for AR displays and proposes design implications for the successful implementation and use of AR in surgical settings.

Index Terms—Augmented Reality, Mixed Reality, HoloLens, Visual Delay, Latency, Surgery, Medicine, Surgical Display

1 Introduction

Augmented reality (AR) involves the superimposition of digital information, such as holograms, onto the real world [4]. One specific type of AR is optical see-through (OST) AR, which allows users to see and interact with both real and virtual objects from a natural first-person perspective. To provide a realistic and immersive experience in AR, it is crucial to maintain the illusion that the virtual and real worlds are part of the same reality. Although several factors can disrupt this illusion, such as low-quality virtual content and limited field of view, latency or visual feedback delay is perhaps the most prominent [42,56]. High latency in AR can occur due to a variety of factors, such as delays in tracking, rendering, and displaying [41]. Long delays between a user's action and the corresponding change in the AR display (often referred to as *'swimming'*) can lead to simulator sickness [51,60] and can impact the user's perception of simultaneity [37,52].

In the real world, the synchronization between what we see and what we do (visuomotor simultaneity) is maintained. However, in the case of AR, the latency introduced by the system can result in delays that can affect motor performance. One solution to high visual feedback delays is the 'move-and-wait' or slowing down approach, where the user moves and then deliberately waits for the system to catch up. However, this approach is not feasible for tasks that require quick and precise movements and have time constraints. For example, during surgery, where AR provides ergonomical benefits [27], a long delay between a surgeon's action and the surgical video feed can lead to fatigue and disorientation [29,68] and potentially endanger patient safety [68] making the surgery impossible to conduct [7].

In general, a low latency is preferred over a high latency. However, it is not always possible to achieve the desired level of latency due to system limitations. This highlights the need to research acceptable latency ranges as this knowledge would aid in designing systems that minimize the risk of latency-related issues. While the human brain is able to tolerate small but perceivable delays and temporarily recalibrates to them [55,70], there is no golden rule specifying the highest acceptable latency threshold for motor tasks. Numerous studies have examined the impact of delays in visual feedback and have investigated the perceptual threshold for delay detection for various motor tasks across different domains (as discussed in Sec. 2), however, they have yielded inconsistent conclusions. For instance, for computer games, some research suggests that users may be unaware of delays below 50 milliseconds (ms) [47], while others suggest that delays as high as 120ms [38], or even 150ms [24] when controlling characters may go unnoticed. Additionally, prior works have shown that the effect of feedback delays for motor tasks is contingent on task predictability [55], task difficulty [24,68], display fidelity [6,66] and user performance [66]. Latency not only just effects what we see but also what we sense, as visual delays can also cause an increase in perceived stiffness [30]. To date, most studies understanding the impact of latency on visuomotor tasks have not been conducted in AR environments and focus mostly

- *Talha Khan, Thomas Downes and, Jacob T. Biehl are with the University of Pittsburgh, School of Computing and Information. E-mail: {muk21, tad85, biehl}@pitt.edu*
- *Toby S. Zhu, Lucille Cheng, Nicolás M. Kass and, Edward G. Andrews are with the University of Pittsburgh Medical Center, School of Medicine. E-mail: {zhuts, chengl2, kassn, andrewse2}@upmc.edu*

on objective performance measures. While the objective outcome is not always affected, latency can have a great impact on users' subjective perception [12]. This gap in literature makes it difficult to compare findings and draw conclusions for acceptable levels of delays for AR applications involving visuomotor tasks.

Our research focuses on the use of AR technology in the field of healthcare, with a particular emphasis on its application in surgical procedures, which has seen a significant growth in recent years [3, 21, 63]. OST AR systems offer a unique advantage in this field by allowing for hands-free interactions, such as voice commands, with holograms while maintaining an unobstructed view of the real world. The use of holograms in the operating room (OR) maintains the integrity of the sterile field and provides unrestricted size and placement affordances. One innovative use of AR in surgery is to use it as a platform for an integrated holographic interface, which presents crucial information directly in front of the surgeon (e.g. patient vital signs, surgical tool video feed, MRI, and more). This can reduce costs and the footprint of the OR by replacing multiple displays with a single AR headset. This holographic interface can also improve the surgeon's visual acuity by allowing them to see and operate on the patient simultaneously, while also using surgical instruments. However, most surgical procedures involve intricate and precise manipulation of surgical tools under tight time constraints and with minimal room for error. A high latency between a surgeon's action and the surgical tool's video feed (e.g. microscope) can negatively impact the surgeon's performance and patient outcomes. Given the demanding nature of surgical procedures, acceptable thresholds of latency for different display modalities in this area can be used as benchmarks for similar fine motor tasks performed in similar settings.

This work aims to investigate the impact of latency for a holographic display in comparison to a conventional display, for motor task performance and subjective perceptions. To achieve this, we conducted a controlled study with a sample of 20 surgeons, where they performed a time-constrained suturing task while observing their actions on either a monitor or a HoloLens 2, under varying levels of latency. The HoloLens 2 is an AR headset which can project virtual content onto the physical world. Therefore, the user has the ability to view the actual physical surroundings and the virtual content simultaneously. Although both display modalities used in the experiment have the ability to display digital content, they have fundamental differences in terms of display fidelity. While a holographic display blends digital content into the real world, a conventional display typically creates a separation between the user and their surroundings. Additionally, for both immersive (VR) and non-immersive environments, the level of display fidelity is known to have effects on task performance, user preference, psychological and physiological reaction and learning [5, 54, 58, 62]. Using our research we seek to answer the following research questions:

1. How does latency in an AR display impact motor task performance (i.e. number of sutures) compared to a conventional display?
2. How does an AR display impact perception (e.g. task load and perceived ranking) of latency compared to a conventional display?

Answers to these questions will provide insights into acceptable latency thresholds for OST AR displays and establish baselines for determining if current technology meets latency requirements for fine motor surgical tasks.

2 Background and Related Work

In the realm of OST AR headsets, several works have made progress in terms of reducing latency. For example, [34] developed an OST AR display capable of achieving an average motion-to-photon (MTP) latency of 80 microseconds (μs). Building on the work of [34], [33] created a high-dynamic-range low-latency AR system with an average MTP latency of 124μs. However, these systems support only 4 degrees of freedom (DoF) and require complex mechanical components. More recently, [45] developed an open-source mixed reality headset that leverages specialized hardware platforms and computer vision algorithms to achieve an average MTP latency of 13.4ms while providing

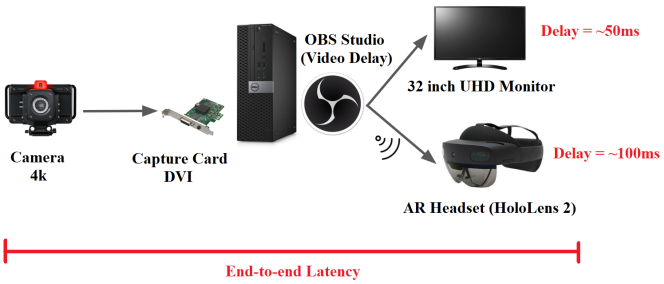

Fig. 2: Pipeline design

6DoF. Unfortunately, these headsets are not currently available for commercial use. It is important to note that these works focus on MTP latency, which is specific to the OST AR headset. In contrast, our work focuses on end-to-end latency, which is the total time it takes for a user's action to be reflected on the display, including the time taken by the camera, capture card, data transmission, and finally the display to complete their respective tasks. These steps are illustrated in Fig. 2.

In our examination of existing research on the impact of latency on human behavior, we identified two key areas of focus: motor task performance and subjective perceptions. The latter consists of latency perception, sense of agency i.e. sense to have caused the action, and, sense of ownership i.e. sense to have undergone the action [19]. We succinctly summarize the findings of previous studies related to our work and contextualize research within this field.

2.1 Impact of latency on motor task performance

Research has shown that delays between performing an action and receiving its visual feedback can negatively impact motor performance. For example, a study conducted by [13] using a see-through haploscope found that higher latencies led to more collisions when operators moved virtual objects along virtual paths. They also report that a latency of less than 50ms is necessary for asymptotic performance. In a simulated AR environment (i.e. an AR environment created in a virtual reality (VR) environment), a similar experimental design to [13] yielded similar trends in results but not exactly the same results as discussed above [32].

Studies based on Fitt's Law [17] style tasks (i.e. well known in HCI research as a predictive model for pointing time) have also found that latency can negatively affect motor performance. For a head-mounted see-through display, [43] found that a higher latency of 146ms more negatively impacted the ability of participants to center a reticle over a moving target. For tapping tasks performed in a VR environment, [59] demonstrated that a high latency increased the time to tap the target. For a 2D mouse-based task, [49] report that target acquisition accuracy drops very quickly for latencies over 50ms. However, tracking error is only slightly affected by latency, with deterioration starting at around 110ms. For a 3D first-person-shooter (FPS) game [23] report that latencies as low as 41ms led to substantial degradation in target acquisition and tracking performance. For pointing and steering tasks, [18] report that delays as low as 16ms can affect motor performance and that the effect is non-linear. Studies that involve full-body avatar-based tasks under varying levels of delays have shown that motor performance is affected by latencies above 75ms [66].

In the surgical domain, studies have consistently shown that higher latencies negatively impact performance, resulting in increased errors, decreased accuracy, and longer task completion times [2, 15, 29, 31, 36, 50, 68]. For instance, a study by [15] found that a 200ms latency led to an 18% increase in task completion time and a 50% increase in errors for a pin transfer task using a robotic arm. Another study by [36] found that a 250ms latency resulted in a 45% increase in task completion time and a 28% increase in tool trajectory length for a block transfer task using robotic hands.

2.2 Impact of latency on perception

Previous research has investigated the just-noticeable difference (JND) - the minimum change in a stimulus that a person can detect - and acceptable latency thresholds in terms of latency perception. Human-computer interaction (HCI) studies report that the JND for video latency in head-mounted displays (HMD) is approximately 15ms [37,48]. While there is a consensus on JND times among studies, the acceptable latency thresholds reported vary widely. For full-body avatar movement, an acceptable threshold of 75ms has been reported [66], while for computer games that require controlling characters, the range is between 50ms-200ms [24, 38, 47, 57]. Additionally, for a task that involved a button press and then required users to observe a visual stimulus with various levels of delay between the two events, users were unable to detect delays below 50ms [52]. The acceptable latency also depends on the input device used [35], with users being able to notice latencies as low as 2ms on touchscreens [44]. Furthermore, latency can increase the perceived difficulty of a task [24], but people can adapt to latency more easily if the task is predictable [55]. Moreover, latency perception also depends on the individual's performance, participants with high-performance errors tend to perceive small delays as non-simultaneous and vice versa [66]. Latency can also affect the sense of presence in virtual environments, where a higher latency results in a reduced sense of presence [40]. In AR applications, [1] have noted that latency degrades the illusion of stability. Telerobotics research in the medical field has reported various acceptable latency thresholds that have decreased over time. For example, an acceptable delay of 700ms was reported by [15] initially but more recent studies suggest it could be as low as 330ms [7] or below 200ms [31]. This highlights the need to study the impact of latency below this threshold. In addition, only a limited number of studies have included usability questionnaires that examine perceived delay, task efficiency, and other factors, as noted in [31, 61]. For instance, research by [31, 62] found that user experience declined when latency exceeded 100ms. However, these studies reported conflicting results regarding whether experienced surgeons were more affected by latency than non-experienced surgeons.

2.3 Distinctions from Prior Work

Based on findings from prior work it is evident that the impact of latency on motor task performance and perception thresholds varies depending on the specific environment (e.g. AR, VR) and the task in question (e.g. predictable/unpredictable, pointing, steering). Therefore, it is important to note that the findings cannot be generalized to other tasks and environments without proper investigation.

In our study, we examine the effect of end-to-end video latency on a fine-motor surgical task in a real-world AR environment and compare it to a traditional monitor. A limitation of previous medical research is its failure to report inherent system latency (e.g. [2,29,36]), this masks the true end-to-end latency experienced by users and thus makes it difficult to draw accurate conclusions.

Moreover, previous studies involving motor tasks performed under varying levels of latency have either been conducted in simulated AR environments (e.g. [32]) or use robotic trainers and simulators (e.g. [31, 68]). In contrast, we conduct our experiments in a real-world AR environment, allowing for realism, intricate hand movements, and haptic feedback - all essential aspects for surgical tasks that are not captured by previous works.

Additionally, simulated AR environments are not equivalent to actual AR environments and can impact experimental results [32]. While most previous studies have focused on objective performance measures, we simultaneously evaluate both motor task performance (objective measure) and subjective measures (e.g. cognitive load) to gain a better understanding of the impact of latency.

Our study focuses on exploring the effects of latency and display modality in surgical AR, which is part of the broader and extensive field of ongoing AR research in the medical domain. The fusion of virtual and physical elements presents particular difficulties in medicine, with notable domain-specific challenges such as instrument tracking and guidance [10, 22, 28], depth perception [11], tele-monitoring [20], and the integration of AI image analysis [9, 10]. Additionally, active efforts have been made to compare the use of AR with traditional methods and interfaces [8].

Table 1: Participant Demographics

			Count
Gender	Male		18
	Female		2
Years of surgical experience		[1-3)	6
		[3-5)	8
		[5-7]	5
		>7	1
Total Participants (N)			20

3 Methods

The study was approved by the university's Institutional Review Board (IRB) and no personally identifiable information was collected. Prior to participation subjects were made aware that this was purely a volunteer opportunity and that no honorarium would be provided upon completion.

3.1 Participants

To determine the study power for our research, we followed established procedures and conducted a within-factors multivariate repeated measures analysis of variance (ANOVA) power analysis using G*Power [16]. Assuming a medium effect size (Cohen's $f = 0.25$), $\alpha = .05$, and 80% power, we aimed for a sample size of 17. In this study, we recruited a total of $N = 20$ participants, including 18 males and 2 females with normal or corrected-to-normal vision. The male-to-female ratio was representative of the gender distribution of surgeons at the hospital where the study was conducted. To ensure that our sample comprised only of experts, we limited our participants to medical residents and attending surgeons, excluding medical students. This decision was based on the recommendation of [31], which suggests that results from non-expert participants may not generalize well for experts. Subjects were recruited from four residency programs (otolaryngology, orthopedic surgery, neurosurgery, urology) at a single university medical center using a combination of word of mouth, phone, and email. All of our participants regularly performed suturing procedures. Nineteen of our participants (medical residents) were in the 26-36 age group, with the exception of a 56-year-old (attending faculty), and the overall median age was 30. Participants had varying levels of years of surgical experience ($\mu = 3.85$, $\sigma = 1.89$) , as detailed in Tab. 1. All participants were in good physical health and regularly participated in surgical procedures at the hospital.

3.2 Equipment

We used a Blackmagic studio camera 4K Pro for capturing video. The camera resolution was down-scaled to 1920 x 1080 and the frame rate was set to 60 frames-per-second (FPS). This was done in order to mimic the display settings surgeons experience in an actual operating room. A Magewell Pro Capture DVI (part number: 11030) was used to capture frames from the camera that were either displayed on a monitor (LG 32inch Ultra HD, model number: 32UN650-W) or the Microsoft HoloLens 2 headset. The study was powered by a Dell Precision 5820 desktop featuring an Intel Xeon W-2223 CPU (3.60 GHz), NVIDIA Quadro RTX 6000 GPU, and 72 GB of RAM.

3.3 End-to-End Latency

We measured end-to-end latency by playing a 60fps millisecond timer video on a display that the camera was facing. We then captured a shot of both the timer display and the monitor frame and HoloLens hologram. The difference in the timers of these displays was used to estimate the end-to-end latency. This method of measuring latency is congruent to [65]. The inherent latency for the monitor setup was 50ms and 100ms for the HoloLens. We performed these experiments several times to validate the end-to-end latency. To add artificial latency we used the video delay filter provided by Open Broadcaster Software

(OBS) Studio. The filter also added an extra latency of 20ms, so for example to add an extra latency of 50ms we only added 30ms to the video delay filter. We performed the latency measurement experiment several times to validate the end-to-end latency for our tasks. It is important to note that all latency measures are approximations because the monitor and HoloLens refresh rates introduce some variability into our measurements, estimated to be around 2-3ms. Therefore, a latency of 50ms fell within the range [47ms-53ms].

3.4 Holographic Streaming

We developed a custom application using Unity 2020 to stream content to the HoloLens. The application consisted of a video player that utilized the onboard computer GPU to render frames. We experimented with several video players in Unity and decided to use the one that introduced the least amount of latency in the pipeline. This was the video player provided by Microsoft MixedReality WebRTC. The application streamed content to the HoloLens using a Holographic Remoting session. During initial prototype development, we tested several communication protocols between the computer and the HoloLens, specifically WebRTC and Microsoft Teams. We found them to be significantly slower than Holographic Remoting and did not use them.

Fig. 1 (right) shows the surgeon's view from a first-person perspective when performing the task using the HoloLens. In order for Holographic Remoting to work, both the HoloLens and the desktop computer were connected to a local network that added a delay of 3-4ms which is included in the inherent latency measurement mentioned above. For the monitor, the camera feed was displayed on OBS Studio as seen in Fig. 1 (middle).

3.5 Experimental Procedure

The study was conducted at a lab located inside an academic hospital. Participants were first informed about the study procedures and objectives and then were asked to sign a consent form and fill out a demographics survey answering questions about age, gender, and years of surgical experience.

Each participant completed seven tasks, four on a monitor with latencies of 50ms, 100ms, 150ms, and 200ms, and three on a HoloLens 2 with latencies of 100ms, 150ms, and 200ms. We selected these latencies for several reasons: the minimum achievable latency for the monitor was 50ms, which is equivalent to actual OR conditions; the minimum achievable latency for the HoloLens was 100ms; latencies above 200ms are unacceptable in surgical tasks, as per prior research [31]; and 50ms intervals represent a delay of three frames per second for a 60fps feed, which is well above the 15ms just noticeable difference (JND) reported by [37, 48]. This allows for better prediction of the true latency experienced by participants. Moreover, data from several levels of latency can be useful for future studies where the minimum achievable latency is close to one of the levels we have tested. For example, if the data has to be modified by a computer vision algorithm, it might also add latency, and our results can help researchers better understand the impact of these added latencies.

The participants performed a suturing task, where the objective was to throw simple uninterrupted stitches using Covidien Surgilon 2-0 braided nylon sutures on a 5 inch x 7 inch (12.7cm x 17.78cm) silicone suture practice pad for three minutes. This task was chosen as suturing is a common surgical procedure and requires precise motor-visual coordination and haptic feedback. These are key components in understanding the effect of latency on surgical procedures. Additionally, the suturing task has also been employed in previous studies to investigate the impact of latency on surgical performance [2, 31, 61, 68].

Before each task, the researchers positioned the suture pad, Adson tweezers, and Olsen Hegar needle driver inside a box, as shown in Fig. 1 (left). The suture pad was marked with points forming a 1cm x 1cm grid. Participants were instructed to keep the lines of sutures inside the grid points. The box had a small opening that allowed participants to insert their hands to perform the necessary maneuvers, simulating a minimally invasive surgical procedure. Participants could only observe their actions on the displays, as the contents of the box were not directly visible. The suture was tied down by a study member prior to the start of each trial so that participants only needed to run the stitch. The timer started when the participant made contact on the suture pad with the suture needle. Each trial terminated when 3 minutes had elapsed, or the participant completed all 16 lines of sutures on the pad, or if an error occurred that prevented task completion, such as detachment of the suture needle. If the total task time was less than 3 minutes, the number of sutures was later adjusted based on extrapolating how many correct sutures the participant could have placed if they had continued for the entire three minutes, rounded to the nearest integer. However, if the error occurred before 90 seconds into the task, it was repeated. Among the total 140 trials only 3 had to be repeated and 25 had to be normalized. After each task, the suture pad was removed, and pictures were taken. A medical student then evaluated two objective measures for each task: the total number of sutures and the number of incorrect sutures. Incorrect sutures were defined as loose sutures, where one loop of the suture was incompletely pulled through and not in contact with the pad, crossing sutures, where two or more lines of suture intersect, or sutures placed outside the marks on the pad. These were defined as errors because in practice, loose, crossing, or misplaced sutures can result in negative consequences for the patient (i.e. wound dehiscence, improper healing, or fistula formation). The order of tasks for each participant was determined using a balanced Latin square design. This yielded the order for the first 14 participants, while the last six participants followed the same order as the first six. Throughout the study participants were unaware of the actual latency they experienced.

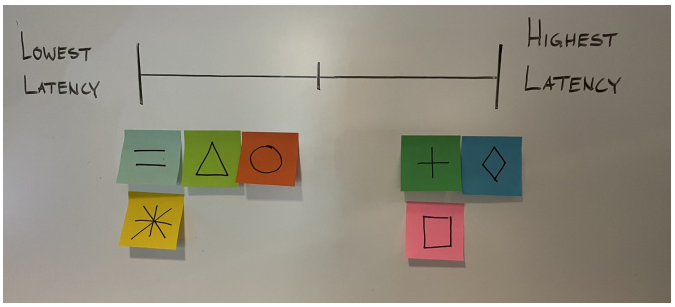

Fig. 3: A subject's ranking of perceived latency

Following the completion of each task, participants were given a sticky note with a unique shape drawn on it, corresponding to a specific latency and display combination. They were then asked to rank their perceived latency by placing the sticky note on a spectrum ranging from low to high, which was drawn on a whiteboard, as shown in Fig. 3. Participants were permitted to adjust their rankings throughout the study session. Subsequently, they completed a survey, which included questions about their perception of latency and its degree of impact on their performance, how the experienced latency would affect an actual surgical procedure in the operating room, a SURG-TLX [67] questionnaire consisting of six factors (mental demands, physical demands, temporal demands, task complexity, situational stress, and distractions), and a simulator sickness questionnaire (SSQ) [25] that measured post-exposure symptoms such as fatigue and dizziness. These measures provide a comprehensive understanding of how participants perceived latency and the impact it had on their performance, enabling us to gain valuable insights into the subjective experiences of the participants during the experiment. A single study session lasted approximately for 90 minutes.

4 Results

In order to assess the performance of surgeons, the study utilized both objective and subjective measures. We present the results for each of these measures separately, beginning with the objective performance evaluation based on the number of sutures placed, followed by the subjective measures, which included perceived latency rankings, the SURG-TLX questionnaire, and the degree of impact of latency on task. In the discussion section (Sec. 5), we will provide explanations for the results obtained from these measures.

To help visualize the data, we have used orange hues for the monitor display and blue hues for the HoloLens display as seen in Fig. 4. Darker colors on the visualizations indicate a higher latency.

All statistical analyses were conducted using either SPSS or the Python *SciPy* library, which are commonly used statistical analysis tools in the research community. Note that for all correlation measurments we performed two-tailed tests for significance.

4.1 Motor Performance

We utilized the number of sutures as a measure of motor performance and conducted a two-way repeated measures ANOVA as it accounts for individual differences among participants. Our data satisfied the assumption of sphericity, as confirmed by Maulchy's test with all $p > 0.05$. Our dependent variable was the number of sutures, while latency, display, and their interaction were the independent variables, and surgical experience (in years) served as the covariate. To maintain a balanced design, we excluded the monitor condition at a latency of 50ms from our analysis. Table 2 presents the means and standard deviations for the number of correct sutures achieved at different display and latency levels. Additionally, our Shapiro-Wilk test verified that our residuals were normally distributed.

We found a significant main effect of display on the number of sutures ($F_{(1,18)} = 7.788$, $p = 0.012$*, $\eta_p^2 = 0.302$), indicating that the type of display used affected motor performance, with participants placing more sutures on average when using the monitor, see Fig. 4 and Fig. 5. We also observed a significant main effect of latency ($F_{(2,36)} = 5.734$, $p = 0.007$*, $\eta_p^2 = 0.242$), suggesting that the amount of delay impacted the number of sutures, with shorter delays resulting in more lines of sutures thrown, see Fig. 4 and Fig. 5. However, we did not find a significant interaction effect for display and latency, suggesting that the effect of display on the number of sutures is consistent across all levels of latency, and vice versa. The experience variable as a covariate also did not have a significant effect on the number of sutures, suggesting it was not a good predictor of the outcome variable. Removing it as a covariate and re-running the analysis resulted in similar observed effects.

For pairwise comparisons between the display types and similar latency levels (e.g., monitor 200ms vs HoloLens 200ms), post-hoc tests using the Bonferroni correction showed no significant difference in performance between the monitor 100ms and the HoloLens 100ms conditions. However, significant differences were observed between both pairs of higher latency conditions. For Pairwise comparisons within the display type across different latency levels (e.g., monitor 150ms vs monitor 200ms), post-hoc results showed a significant impact on performance only between the 100ms and 200ms conditions for both display types but non-significant for all other conditions. Tab. 3 reports relevant measures for the post-hoc tests.

To examine differences between the monitor 50ms, monitor 100ms, and HoloLens 100ms conditions, we conducted two one-way repeated measures ANOVA tests. The Levene's test for homogeneity of variance yielded a non-significant result, indicating that equality of variance was preserved across conditions. However, we did not find a significant difference among any of the tested conditions. Specifically, the effect size for the comparison between the 50ms monitor and 100ms monitor conditions was $\eta_p^2 = 0.016$, and the effect size for the comparison between the 50ms monitor and 100ms HoloLens conditions was $\eta_p^2 = 0.065$. These values of effect sizes can be taken as further evidence of the null effect.

4.1.1 Speed and Accuracy Tradeoff

We performed a Spearman rank correlation analysis to examine the relationship between the speed of suturing and the accuracy of sutures, specifically the number of correct sutures achieved. For the monitor display, we found a significant moderate negative correlation between latency and the number of correct sutures, with a correlation coefficient of $r = -0.389$ and $p < 0.001$*. However, we found a negligible positive and non-significant correlation between latency and the number of incorrect sutures (i.e., $p > 0.05$). Similarly, for the HoloLens display, we found a significant moderate negative correlation between latency and the number of correct sutures, with a correlation coefficient of $r = -0.347$ and $p < 0.001$*. However, we found a negligible negative and non-significant correlation between latency and the number of incorrect sutures (i.e., $p > 0.05$). These findings suggest that as the delay in display increased, the number of correct sutures decreased, while the number of incorrect sutures remained unaffected for both display types (refer to Fig. 5). Therefore, it appears that the participants did not compromise accuracy for speed.

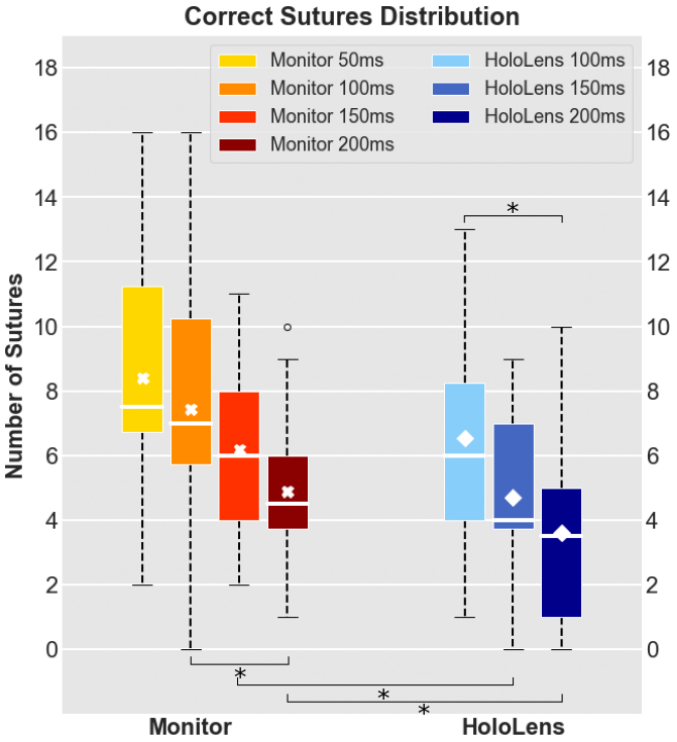

Fig. 4: Distribution of all conditions for the number of correct sutures. The cross and diamond shapes inside the plot represent the mean of the condition

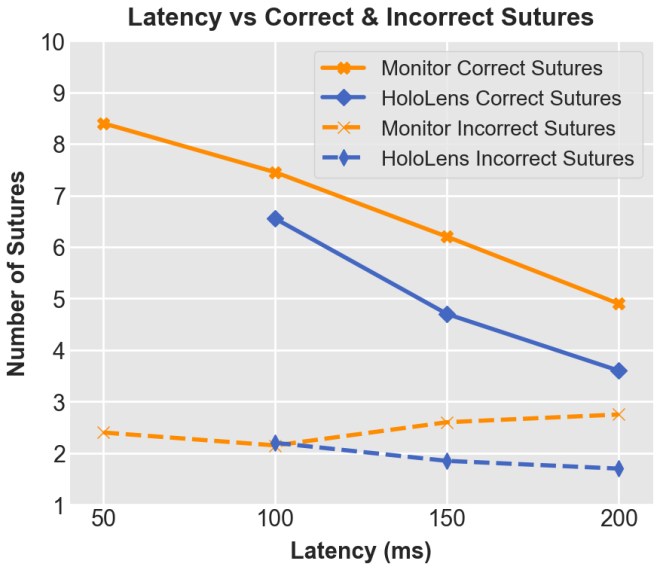

Fig. 5: Impact of latency on the mean number of correct and incorrect sutures for both displays

Table 2: Means and standard deviations $\mu(\sigma)$ of motor performance and SURG-TLX scores

Measure \ Latency	50 M	100 M	150 M	200 M	100 HL	150 HL	200 HL
Correct Number of Sutures	8.40(3.63)	7.45(3.96)	6.20(2.70)	4.90(2.47)	6.55(3.56)	4.70(2.63)	3.60(2.95)
SURG-TLX Score	124.0(108.0)	124.0(112.0)	137.35(121.48)	143.75(115.19)	123.55(108.55)	150.60(117.32)	182.65(124.48)

Table 3: Pairwise comparisons for latency and display for number of sutures, adjusted for multiple comparisons (with Bonferroni correction).

		Number of Sutures		
	Latency (ms)	Mean Dif.	Std. Error	Sig.
Monitor vs HoloLens	100 - 100	0.90	0.692	>0.05
	150 - 150	1.50	0.675	0.039*
	200 - 200	1.30	0.559	0.032*
Monitor vs Monitor	100 - 150	1.25	0.721	>0.05
	150 - 200	1.30	0.568	>0.05
	100 - 200	2.55	0.723	0.007*
HoloLens vs HoloLens	100 - 150	1.85	0.733	>0.05
	150 - 200	1.10	0.619	>0.05
	100 - 200	2.95	0.781	0.004*

4.2 Latency Perception

To analyze the perceived latency, we utilized participants' responses to **(1)** the SURG-TLX questionnaire, which included six factors: mental demands, physical demands, temporal demands, task complexity, situational stress, and distractions. **(2)** The question regarding the presence of latency and the degree of impact on the task, with response options listed in Fig. 8. **(3)** Perceived latency rankings, where participants were asked to place a unique sticky code on a board with a spectrum of latency ranging from low to high, as shown in Fig. 3.

4.2.1 SURG-TLX

The SURG-TLX was summed across all factors to yield a single score for each task. This is consistent with prior studies that have also used the SURG-TLX [26]. We had to $log(x+1)$ transform our data in order to ensure that our residuals were normally distributed. All statistical tests were conducted using the transformed data, but the figures and tables represent the untransformed data. In general, as latency increased the task load increased and we observed a steeper increase for the HoloLens than the monitor, refer to Fig. 7. In Table 2, we present the means and standard deviations for the summed SURG-TLX scores at different display and latency levels.

We then ran a two-way repeated measures ANOVA. Our dependent variable was the SURG-TLX score, while latency, display, and their interaction were the independent variables, and surgical experience (in years) served as the covariate. Our data satisfied the assumption of sphericity, as confirmed by Maulchy's test with all $p > 0.05$. We found no significant effects for the display, latency, and their interaction, all $p > 0.05$. However, we did find a significant effect for experience. To better understand this relationship we computed a Pearson's rank correlation for both the monitor and HoloLens displays between the experience and SURG-TLX score. We found a significant negative correlation for both the monitor ($r(80) = -0.396$, $p < 0.001$*) and HoloLens ($r(60) = -0.361$, $p < 0.001$*), indicating that more experienced surgeons experienced less task load.

To test for differences between the monitor 50ms, monitor 100ms, and HoloLens 100ms conditions, we conducted a one-way ANOVA. The Levene's test for homogeneity of variance yielded a non-significant result, indicating that equality of variance was preserved across conditions. However, we did not find a significant difference among any of the tested conditions. Specifically, the effect size for the comparison between the 50ms monitor and 100ms monitor conditions was η_p^2 =0.006, and the effect size for the comparison between the 50ms monitor and 100ms HoloLens conditions was η_p^2 =0.002. These values of effect sizes can be taken as further evidence of the null effect.

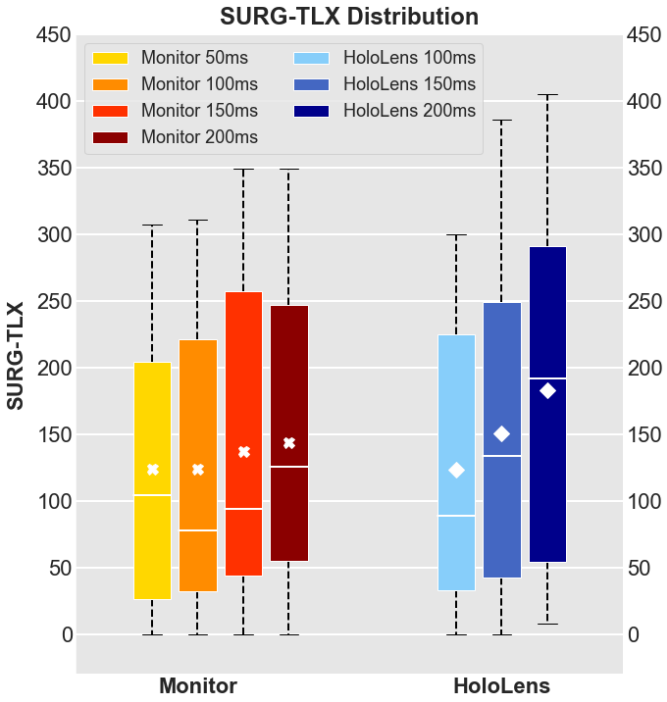

Fig. 6: Distribution of all conditions for the SURG-TLX scores. The cross and diamond shapes inside the plot represent the mean of the conditions

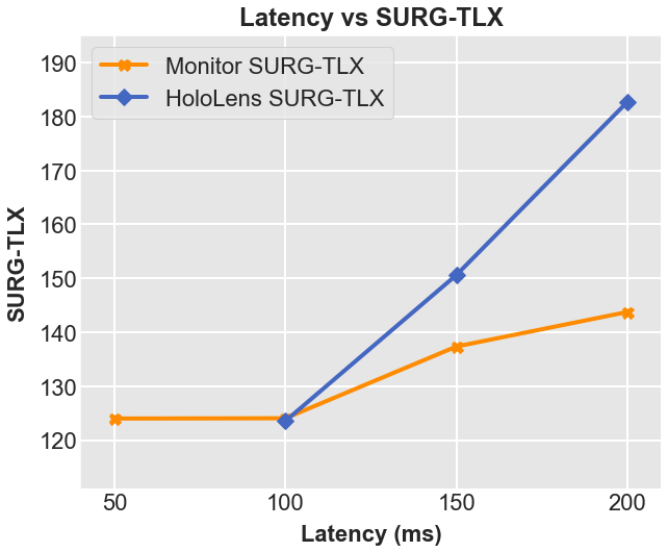

Fig. 7: Impact of latency on the mean SURG-TLX score

Table 4: Means and standard deviations of responses for all conditions for the question on latency presence and its degree of impact on task (M = monitor, H = HoloLens). Note that **1** = was not present, **2** was present, no impact, **3** = was present, mild impact, **4** was present, moderate impact, **5** was present, significant impact

	M 50ms	M 100ms	H 100ms	M 150ms	H 150ms	M 200ms	H 200ms
μ	2.20	2.45	2.30	2.90	3.35	3.15	3.95
σ	1.21	1.24	0.90	1.13	1.15	0.85	1.02

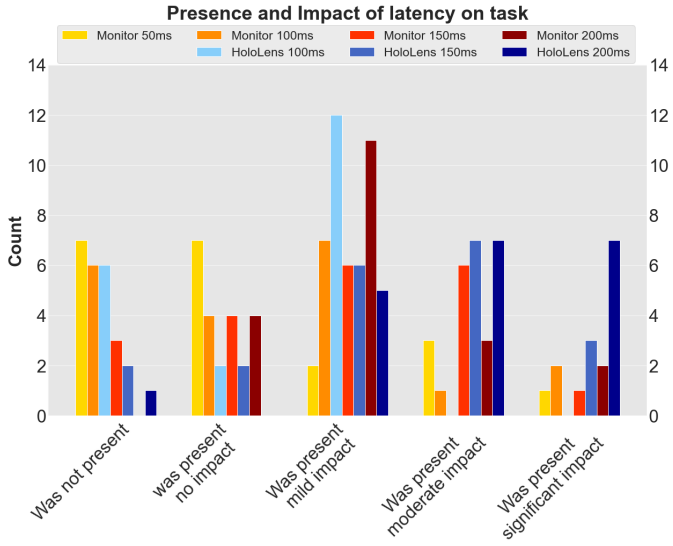

Fig. 8: Count of responses for survey question on presence of latency and its impact on task

4.2.2 Latency Presence and Impact on Task

The participants were asked to respond to a question after completing each task, inquiring whether they perceived any latency and the extent of its impact on the task. Participants were provided with five options (shown in Fig. 8) and their responses were subsequently coded to a numerical scale of 1 to 5, with 1 indicating the absence of latency and 5 representing significant impact due to latency. The average rankings for each task are presented in Tab. 4.

Overall the findings suggest that as the latency increased, the perception of its presence and impact on the task also increased. For the lower latency pair (monitor 100ms, HoloLens 100ms), the participants, on average, reported feeling less impact on the task for the HoloLens. However, for the higher latency pairs (150ms and 200ms), participants, on average, reported feeling more impact on the task for the HoloLens than the monitor. To further investigate the relationship between latency and its subjective impact on the task, we computed the Pearson rank correlation for both the monitor and the HoloLens. The results indicated a moderate positive significant correlation for the monitor, with $r(80) = 0.312$, $p = 0.005$*, and a strong positive significant correlation for the HoloLens, with $r(60) = 0.545$, $p < 0.001$*. This indicates that users perceived a higher latency on the HoloLens to have a greater impact on the task compared to the monitor.

4.2.3 Perceived Latency Rankings

Upon completing each task, the participants were asked to rank the perceived latency on a spectrum using sticky notes, ranging from low to high latency as shown in Fig. 3. To convert these rankings into numerical values, we assigned a rank to each latency based on the position of the corresponding sticky note on the spectrum. In cases where multiple sticky notes were stacked on the spectrum, all tasks associated with those notes were assigned the same rank, and subsequent ranks were skipped. Tab. 5 presents the descriptive statistics for these rankings,

Table 5: Subjective rankings of actual latencies perceived by subjects

	Rankings								
	1	**2**	**3**	**4**	**5**	**6**	**7**	μ	σ
Monitor 50ms	12	3	2	0	1	0	1	1.89	1.62
Monitor 100ms	4	8	0	2	2	3	0	2.94	1.84
HoloLens 100ms	6	3	5	4	1	0	0	2.52	1.31
Monitor 150ms	1	3	5	6	0	2	2	3.79	1.65
HoloLens 150ms	0	2	3	2	3	5	4	4.94	1.71
Monitor 200ms	1	2	1	5	6	3	1	4.36	1.53
HoloLens 200ms	0	0	0	1	5	4	9	6.11	0.99

excluding subject 10's data as they misunderstood the ranking task and provided pairwise comparisons rather than overall comparisons.

The results indicated that for a latency of 100ms, on average, the participants perceived the HoloLens display to have shorter latency than the monitor. However, for higher latencies (150ms and 200ms), the participants ranked the monitor latency lower than the HoloLens. To further explore the relationship between actual latency and perceived latency, we computed the Pearson rank correlation coefficients for both the monitor and the HoloLens. The results showed a moderate positive significant correlation for the monitor, with $r(76) = 0.491$, $p < 0.001$, and a very strong positive significant correlation for the HoloLens, with $r(57) = 0.730$, $p < 0.001$. These findings suggest that as latency increased, perceived latency also increased. However, the correlation coefficient was higher for the HoloLens, indicating that users are more sensitive to latency on the HoloLens compared to the monitor.

4.3 Performance and Perceived Latency Relationship

We conducted an analysis to investigate the impact of performance (i.e., the number of sutures) on the perception of latency. Specifically, we performed a multiple regression analysis with actual latency, number of sutures, and their interaction as independent variables, and the perceived latency rankings provided by the study participants as the dependent variable.

For the monitor, the model showed a marginally good fit for the data, with an R-squared value of 0.25 ($F_{(3,72)} = 7.981$, $p < 0.001$*). The results indicated that latency was a marginally significant predictor of perceived latency ($\beta = 0.014$, $t = 1.739$, $p = 0.086$). However, the number of sutures and the interaction of latency and the number of sutures were non-significant predictors of perceived latency.

For the HoloLens, the model was a good fit for the data, with an R-squared value of 0.559 ($F_{(3,53)} = 22.386$, $p < 0.001$*). The results revealed that latency was a significant predictor of perceived latency ($\beta = 0.037$, $t = 4.612$, $p < 0.001$*). However, similar to the monitor case, both the number of sutures and the interaction were non-significant predictors of perceived latency. We confirmed that the assumptions of normality and linearity were met for both models.

These findings suggest that participants relied more on the actual latency to rank the perceived latency rather than the number of sutures (their motor performance). Figure 9 displays the relationship between actual latency, perceived latency, and the number of sutures.

4.4 Feasibility in the OR

After completing each task, participants were asked to answer a question regarding how the experienced latency would affect an actual surgical procedure in the operating room. They were allowed to select multiple options from a list of possibilities, as shown in Fig. 10. The results indicate that as latency increased, participants suggested that errors, completion time, and risk of patient safety would also increase.

To further investigate, we conducted a chi-square test to examine whether there were any differences in the counts of participant responses between the HoloLens and monitor for any of the response options, for all latency conditions. However, no significant differences were found, all p-values > 0.05.

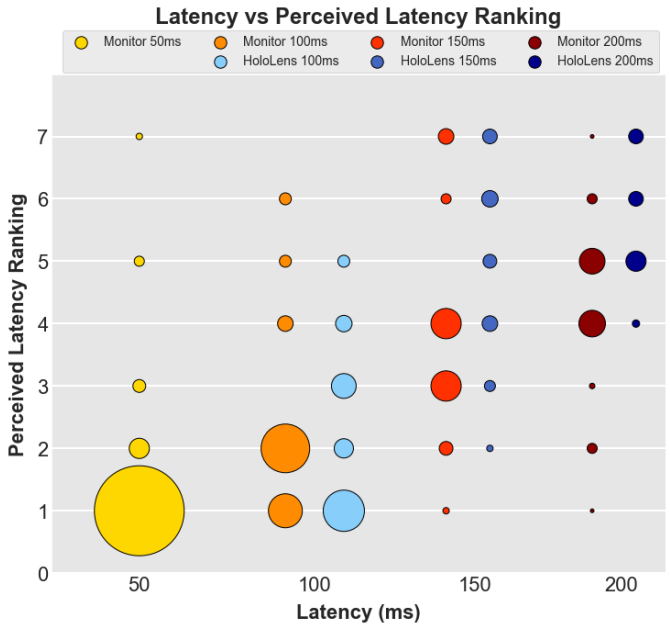

Fig. 9: Relationship between actual latency, perceived latency rankings and the number of sutures. The area of the circles represents the number of sutures. If multiple users reported the same perceived latency, the number of sutures used was summed across all those participants.

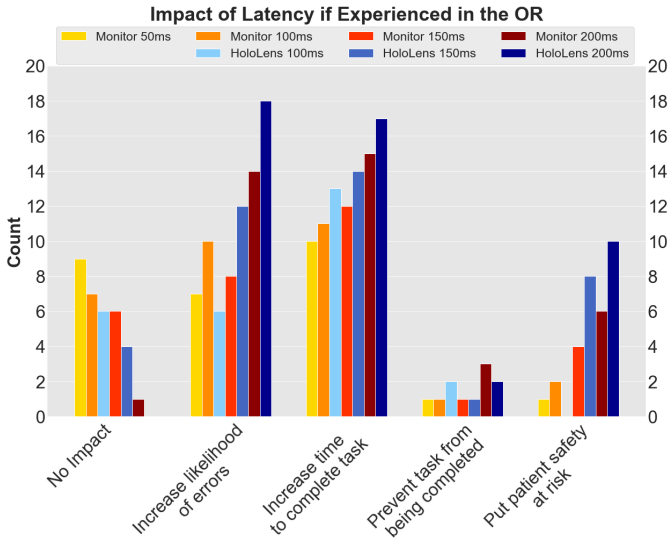

Fig. 10: Count of responses for survey question on the impact of latency if experienced during an actual surgical procedure in the operating room.

4.5 Physical Well-being

After completing each task, participants were asked to respond to the Simulator Sickness Questionnaire (SSQ), which contained questions about their experience with motion sickness, headache, fatigue, and other related symptoms. Participants were asked to select one of four response options for each category: *'not at all'*, *'mild'*, *'moderate'*, *and 'severe'*.

Overall, the results were largely positive, with a vast majority of participants selecting the option *'not at all'* for all categories except for *'Eyestrain'*, where around half of the responses were *'not at all'* and the rest were *'mild'*. Importantly, there was an even distribution of these responses across both display modalities and latency levels, suggesting that the eyestrain symptoms were independent of these factors.

These findings are consistent with a prior study [64], which investigated whether the Microsoft HoloLens caused simulator sickness in a sample of 142 participants, including 46 medical professionals. The study found that the HoloLens caused only negligible symptoms of sickness, with most users reporting no symptoms and only a small minority reporting general discomfort.

5 Discussion

This section is structured into three parts: first, we reflect on our research questions and discuss the implications of our findings; second, we discuss the acceptability of using the HoloLens in surgical procedures; and third, we propose a solution to address the current limitations arising from high latency in surgical systems.

Our findings are consistent with previous research, both in the HCI and medicine domains, which has demonstrated that as latency increases, task performance declines [2, 15, 36, 43, 59]. In our analysis, we found that both the level of latency and the display type significantly impacted motor task performance. For the 100ms condition we found no significant difference between the HoloLens and the monitor, with a small effect size (Cohen's $d = 0.238$), suggesting comparable performance for both display modalities. However, for the 150ms and 200ms latency conditions, we found a significant difference in performance between the monitor and the HoloLens, with medium effect sizes (Cohen's $d = 0.561$ and 0.478, respectively). Participants performed better using the monitor for both of these latency levels. This suggests that at higher latencies, performance deteriorates more when using the HoloLens compared to the monitor. We speculate this could be due to the surgeons' having less experience using the HoloLens compared to the monitor. Further studies are needed to investigate whether comparable performance can be achieved on both display modalities at higher latency levels with normalized experience.

Interestingly, we did not observe any effect of surgeons' experience on task performance. This may be due to the fact that our participants were all experts who had likely reached asymptotic performance levels for the suturing task through extensive training. However, medical students or non-experts may exhibit different patterns of task performance with changes in latency and display, as shown in [31].

Our study did not reveal any evidence of participants compromising accuracy for speed. Regardless of latency level, participants prioritized accuracy by adjusting their task speed when faced with higher latencies. We speculate this behavior can be attributed to surgical training that prioritizes accuracy over speed to ensure patient safety. However, non-experts may exhibit different behavior, as demonstrated in [13], where subjects were observed to compromise accuracy for speed. **Therefore, we believe future systems should incorporate feedback mechanisms to inform users when the latency is high and prompt them to adjust their behavior to avoid errors. This will make future systems more inclusive and resilient to different levels of user experience.**

With regard to latency perception, our study investigated the relationship between latency, task load, and experience, as measured by the SURG-TLX. We found that as latency increased, participants perceived an increase in task load. While we did not observe any significant effects with display type or latency level on task load, we did observe a significant effect with training level, as more experienced surgeons had lower task loads even at higher latencies. This result is consistent with previous studies, such as [61]. Taken together, these results suggest that cognitive load is inherent to the participant and likely specific to the task. **This finding may have important implications for the design of future systems, as reducing latency alone may not necessarily lead to a decrease in task load, especially among less experienced users. Therefore, future studies could investigate additional factors that may affect task load, such as task difficulty, interface design, and more.**

As latency increased, participants were able to detect the change and indicated its severe impact on the task. On average, participants found that higher latencies on the HoloLens (150ms and 200ms) had a greater impact on task performance compared to the 100ms latency condition. In general participants' perception of perceived latency was representa-

tive of the true latency they encountered. One possible explanation is that the just-noticeable difference (JND) for headsets is 15ms [37,48], and since our conditions had a difference of 50ms participants were able to provide correct rankings. However, on average, participants were more sensitive to higher latencies (150ms and 200ms conditions) on the HoloLens compared to the monitor. Despite surgeons' motor performance being lower on the HoloLens for the 100ms condition, on average, they perceived a lower latency and lower impact for this condition. We speculate this is due to the ergonomic affordances provided by the HoloLens, where participants were able to place the hologram at a more comfortable working position.

The study found that participants' perception of latency was not affected by their performance, which conflicts with a previous study conducted by [66]. The previous study reported that participants relied more on their own performance than the actual delay when estimating simultaneity feedback. For instance, in that study, if a person did not perform well at a lower latency, they would perceive it as high. The discrepancy in results may be attributed to differences in subject pools, as our participants were experts in their field, whereas [66] participants were not. **This finding also highlights the importance of providing latency feedback, especially for novice users who may be more affected by latency than experts. By letting them know that their performance is causing the issue and not the delay in the system, they may be able to improve their performance through behavior modification.**

Our study results demonstrate that participants were able to detect latency in the 50ms monitor condition, as depicted in Fig. 8. This is similar to previous findings in [52] where gamers could detect delays as low as 50ms. It is important to note that 50ms of latency is present in state-of-the-art operating rooms displays (measured at a major academic US hospital), yet surgeons are still able to conduct surgeries with this level of latency. This may be due to latency training, where with enough time and experience with a particular latency level, surgeons can learn to improve their performance [69]. However, this raises the question of what is the threshold of acceptable latency. Previous studies, discussed in Sec. 2, have either directly asked participants whether the latency was acceptable or measured deterioration in performance to estimate an acceptable threshold. Our approach is different; we compared a known acceptable threshold of 50ms with higher latency levels. Interestingly, we found no significant difference in motor performance or surgical task load between the 50ms monitor condition and the 100ms conditions for both the monitor and the HoloLens. However, we did observe a significant difference between the 100ms and 200ms conditions for both devices, suggesting that the motor performance deteriorates significantly at the 200ms mark, and an acceptable latency threshold must be below 200ms, as suggested in [31]. Examining Fig. 10, we see that the latency of 150ms also cannot be considered acceptable, as eight surgeons using the HoloLens suggested that it could potentially endanger patient safety, compared to zero surgeons reporting safety concerns in the 100ms condition.

However, our results alone cannot argue that the HoloLens with a latency of 100ms is acceptable for intraoperative use with actual patients. Future studies need to investigate whether the HoloLens can sustain surgical tasks lasting several hours, which typically last 3-7 hours [39], and effects of this cognitive load on users over longer operations. Moreover, our experiments were conducted on a predictable and specific task in a controlled environment, and future studies must account for this limitation. **Nevertheless, our finding of no significant difference between the 50ms and 100ms conditions provides a foundation for seeking regulatory approval (e.g., FDA clearance in the US) for clinical trials using the HoloLens and moves us one step closer to incorporating augmented reality inside the operating room.**

Achieving zero latency in a system would be the ideal solution. However, we recognize that this is likely not feasible in practice due to physical limitations of current electronic systems (e.g., even light does not travel instantaneously). Nevertheless, there are ways to mitigate the effects of latency by improving hardware or devising efficient software solutions. Regarding hardware, commercially developing AR headsets with low latency [33,34,45] would be the most straightforward solution to address high latency. However, even with low-latency headsets, there could be additional delays introduced by other steps in the processing pipeline, for example, image processing to detect tumors [14] or porting multiple data streams (e.g. patient vital signs, scans, and endoscopic video feeds). In the software domain, predictive AR has been proposed as a solution to alleviate the effects of latency [53]. This approach has been shown to decrease task completion time without affecting error rates, but its feasibility in the OR needs further evaluation, and subjective data need to be collected. Motion scaling has also been proposed to reduce errors in high-latency systems, but has only be proven to work in robot-assisted environments [46]. Faster image generation approaches have also been suggested [71] but it is unsure if they are implemented in commercially available AR headsets. **A more contemporary solution would be to use a hybrid setup that combines an AR display with a backup monitor display. When the latency in the AR display reaches a certain threshold, the system could notify the user to stop or switch to the monitor display with lower latency. These thresholds don't have to be strict but rather can be dynamic depending on the surgeon, for example, more experienced surgeons can manually set higher thresholds than less experienced ones. This approach can ensure that surgeons always have access to information while minimizing the effects of latency on surgical performance.**

6 Limitations and Future Work

We acknowledge several limitations in our study. First, there was a gender imbalance among our participants. Second, the duration of the task was limited, and it is unclear how our findings would generalize to longer surgeries in real-world settings. Third, our sample size was insufficient to perform a detailed analysis of self-reported data.

To address the limitations identified in our research, several recommendations can be made for future studies. Firstly, we propose conducting experiments over an extended period to allow participants to become more familiar with and accustomed to using the HoloLens. This extended exposure may lead to improved performance and more accurate perceptions of the technology. Additionally, incorporating qualitative methods, such as interviews, can provide valuable insights into the observed differences between different display modalities. By gaining a deeper understanding of participants' perspectives and experiences, we can better interpret the quantitative data and identify potential areas for improvement. To enhance the comprehensiveness of our findings, it is advisable to collect physiological data alongside self-reported measures. This can provide a more holistic understanding of participants' perceptions. Furthermore, exploring the use of pass-through VR headsets in future studies can offer better control over what participants see. However, it is important to consider the delay introduced to both real-world and virtual objects when employing such headsets, as it may impact user performance and overall experience. By following these recommendations, we can overcome limitations and gain deeper insights into the effectiveness and usability of the HoloLens in surgical settings. The studies by [40] on collecting physiological data and [42] on the delay introduced by pass-through VR headsets provide valuable supporting evidence for these suggestions.

7 Conclusion

We conducted a novel study to investigate the impact of various levels of latency on the use of AR for a fine motor surgical task. We presented quantitative comparisons for both objective and subjective assessment measures for the monitor and HoloLens 2. Our findings indicate that both level of latency and display type impact motor performance, and higher latencies on the HoloLens result in relatively poor performance. However, cognitive load was found to be unrelated to display type or latency, but was dependent on the surgeon's training level (experience). Surgeons did not compromise accuracy to gain more speed and were generally well aware of the latency in the system irrespective of their performance on task. Based on our findings, we suggest a suitable latency threshold for the HoloLens 2 when used for surgical tasks. Finally, our study offers valuable design implications for mitigating the impact of latency in future AR systems developed for fine motor tasks.

REFERENCES

[1] R. S. Allison, L. R. Harris, M. Jenkin, U. Jasiobedzka, and J. E. Zacher. Tolerance of temporal delay in virtual environments. In *Proceedings IEEE Virtual Reality 2001*, pp. 247–254. IEEE, 2001. 3

[2] M. Anvari, T. Broderick, H. Stein, T. Chapman, M. Ghodoussi, D. W. Birch, C. Mckinley, P. Trudeau, S. Dutta, and C. H. Goldsmith. The impact of latency on surgical precision and task completion during robotic-assisted remote telepresence surgery. *Computer Aided Surgery*, 10(2):93–99, 2005. 2, 3, 4, 8

[3] A. Ayoub and Y. Pulijala. The application of virtual reality and augmented reality in oral & maxillofacial surgery. *BMC Oral Health*, 19:1–8, 2019. 2

[4] R. T. Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997. 1

[5] W. Barfield, C. Hendrix, and K. Bystrom. Visualizing the structure of virtual objects using head tracked stereoscopic displays. In *Proceedings of IEEE 1997 Annual International Symposium on Virtual Reality*, pp. 114–120. IEEE, 1997. 2

[6] D. A. Bowman, C. Stinson, E. D. Ragan, S. Scerbo, T. Höllerer, C. Lee, R. P. McMahan, and R. Kopper. Evaluating effectiveness in virtual environments with mr simulation. In *Interservice/Industry Training, Simulation, and Education Conference*, vol. 4, p. 44, 2012. 1

[7] S. E. Butner and M. Ghodoussi. Transforming a surgical robot for human telesurgery. *IEEE Transactions on Robotics and Automation*, 19(5):818–824, 2003. 1, 3

[8] F. M. Calisto, A. Ferreira, J. C. Nascimento, and D. Gonçalves. Towards touch-based medical image diagnosis annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, pp. 390–395, 2017. 3

[9] F. M. Calisto, N. Nunes, and J. C. Nascimento. Breastscreening: on the use of multi-modality in medical imaging diagnosis. In *Proceedings of the international conference on advanced visual interfaces*, pp. 1–5, 2020. 3

[10] E. Castelan, M. Vinnikov, and X. Alex Zhou. Augmented reality anatomy visualization for surgery assistance with hololens: Ar surgery assistance with hololens. In *ACM International Conference on Interactive Media Experiences*, pp. 329–331, 2021. 3

[11] N. H. Christensen, O. G. Hjermitslev, F. Falk, M. B. Madsen, F. H. Østergaard, M. Kibsgaard, M. Kraus, J. Poulsen, and J. Petersson. Depth cues in augmented reality for training of robot-assisted minimally invasive surgery. In *Proceedings of the 21st International Academic Mindtrek Conference*, pp. 120–126, 2017. 3

[12] M. Dick, O. Wellnitz, and L. Wolf. Analysis of factors affecting players' performance and perception in multiplayer games. In *Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games*, pp. 1–7, 2005. 2

[13] S. R. Ellis, F. Breant, B. Manges, R. Jacoby, and B. D. Adelstein. Factors influencing operator interaction with virtual objects viewed via head-mounted see-through displays: viewing conditions and rendering latency. In *Proceedings of IEEE 1997 Annual International Symposium on Virtual Reality*, pp. 138–145. IEEE, 1997. 2, 8

[14] Engadget. Nvidia and medtronic are building an ai-enhanced endoscopy tool. https://www.engadget.com/nvidia-and-medtronic-are-building-an-ai-enhanced-endoscopy-tool-161532723.html, October 2022. 9

[15] M. D. Fabrlzio, B. R. Lee, D. Y. Chan, D. Stoianovici, T. W. Jarrett, C. Yang, and L. R. Kavoussi. Effect of time delay on surgical performance during telesurgical manipulation. *Journal of endourology*, 14(2):133–138, 2000. 2, 3, 8

[16] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007. 3

[17] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology: General*, 121(3):262, 1992. 2

[18] S. Friston, P. Karlström, and A. Steed. The effects of low latency on pointing and steering tasks. *IEEE transactions on visualization and computer graphics*, 22(5):1605–1615, 2015. 2

[19] S. Gallagher. Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences*, 4(1):14–21, 2000. 2

[20] D. Gasques, J. G. Johnson, T. Sharkey, Y. Feng, R. Wang, Z. R. Xu, E. Zavala, Y. Zhang, W. Xie, X. Zhang, et al. Artemis: A collaborative mixed-reality system for immersive surgical telementoring. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021. 3

[21] H. Ghaednia, M. S. Fourman, A. Lans, K. Detels, H. Dijkstra, S. Lloyd, A. Sweeney, J. H. Oosterhoff, and J. H. Schwab. Augmented and virtual reality in spine surgery, current applications and future potentials. *The Spine Journal*, 21(10):1617–1625, 2021. 2

[22] C. Gsaxner, J. Li, A. Pepe, D. Schmalstieg, and J. Egger. Inside-out instrument tracking for surgical navigation in augmented reality. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11, 2021. 3

[23] Z. Ivkovic, I. Stavness, C. Gutwin, and S. Sutcliffe. Quantifying and mitigating the negative effects of local latencies on aiming in 3d shooter games. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pp. 135–144, 2015. 2

[24] S. Jörg, A. Normoyle, and A. Safonova. How responsiveness affects players' perception in digital games. In *Proceedings of the ACM symposium on applied perception*, pp. 33–38, 2012. 1, 3

[25] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. 4

[26] L. R. Kennedy-Metz, H. L. Wolfe, R. D. Dias, S. J. Yule, and M. A. Zenati. Surgery task load index in cardiac surgery: measuring cognitive load among teams. *Surgical innovation*, 27(6):602–607, 2020. 6

[27] T. Khan, E. G. Andrews, P. A. Gardner, A. N. Mallela, J. R. Head, J. C. Maroon, G. A. Zenonos, D. Babichenko, and J. T. Biehl. Ar in the or: exploring use of augmented reality to support endoscopic surgery. In *ACM International Conference on Interactive Media Experiences*, pp. 267–270, 2022. 1

[28] T. Khan, J. T. Biehl, E. G. Andrews, and D. Babichenko. A systematic comparison of the accuracy of monocular rgb tracking and lidar for neuronavigation. *Healthcare Technology Letters*, 2022. 3

[29] T. Kim, P. Zimmerman, M. Wade, and C. Weiss. The effect of delayed visual feedback on telerobotic surgery. *Surgical Endoscopy and Other Interventional Techniques*, 19(5):683–686, 2005. 1, 2, 3

[30] B. Knorlein, M. Di Luca, and M. Harders. Influence of visual and haptic delays on stiffness perception in augmented reality. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pp. 49–52. IEEE, 2009. 1

[31] A. Kumcu, L. Vermeulen, S. A. Elprama, P. Duysburgh, L. Platiša, Y. Van Nieuwenhove, N. Van De Winkel, A. Jacobs, J. Van Looy, and W. Philips. Effect of video lag on laparoscopic surgery: correlation between performance and usability at low latencies. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 13(2):e1758, 2017. 2, 3, 4, 8, 9

[32] C. Lee, S. Bonebrake, D. A. Bowman, and T. Höllerer. The role of latency in the validity of ar simulation. In *2010 IEEE Virtual Reality Conference (VR)*, pp. 11–18. IEEE, 2010. 2, 3

[33] P. Lincoln, A. Blate, M. Singh, A. State, M. C. Whitton, T. Whitted, and H. Fuchs. Scene-adaptive high dynamic range display for low latency augmented reality. In *Proceedings of the 21st ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 1–7, 2017. 2, 9

[34] P. Lincoln, A. Blate, M. Singh, T. Whitted, A. State, A. Lastra, and H. Fuchs. From motion to photons in 80 microseconds: Towards minimal latency for virtual and augmented reality. *IEEE transactions on visualization and computer graphics*, 22(4):1367–1376, 2016. 2, 9

[35] M. Long and C. Gutwin. Effects of local latency on game pointing devices and game pointing tasks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019. 3

[36] M. J. Lum, J. Rosen, H. King, D. C. Friedman, T. S. Lendvay, A. S. Wright, M. N. Sinanan, and B. Hannaford. Teleoperation in surgical robotics–network latency effects on surgical performance. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6860–6863. IEEE, 2009. 2, 3, 8

[37] K. Mania, B. D. Adelstein, S. R. Ellis, and M. I. Hill. Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, pp. 39–47, 2004. 1, 3, 4, 9

[38] M. Mauve, J. Vogel, V. Hilt, and W. Effelsberg. Local-lag and timewarp: Providing consistency for replicated continuous applications. *IEEE transactions on Multimedia*, 6(1):47–57, 2004. 1, 3

[39] P. Medicine. Brain tumor surgery faq. https://www.pennmedicine.org/for-patients-and-visitors/find-a-program-or-service/neurosurgery/brain-tumor-center/

brain-tumor-surgery-faq, Accessed 2023-03-23. 9

[40] M. Meehan, S. Razzaque, M. C. Whitton, and F. P. Brooks. Effect of latency on presence in stressful virtual environments. In *IEEE Virtual Reality, 2003. Proceedings.*, pp. 141–148. IEEE, 2003. 3, 9

[41] M. R. Mine. Characterization of end-to-end delays in head-mounted display systems. *The University of North Carolina at Chapel Hill, TR93-001*, 1993. 1

[42] M. Nabiyouni, S. Scerbo, D. A. Bowman, and T. Höllerer. Relative effects of real-world and virtual-world latency on an augmented reality training task: an ar simulation experiment. *Frontiers in ICT*, 3:34, 2017. 1, 9

[43] W. T. Nelson, R. S. Bolia, C. A. Russell, R. M. Morley, and M. M. Roe. Head-slaved tracking in a see-through hmd: The effects of a secondary visual monitoring task on performance and workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, pp. 3–390. SAGE Publications Sage CA: Los Angeles, CA, 2000. 2, 8

[44] A. Ng, J. Lepinski, D. Wigdor, S. Sanders, and P. Dietz. Designing for low-latency direct-touch input. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pp. 453–464, 2012. 3

[45] T. Nguyen. Low-latency mixed reality headset. *Low-latency VR/AR Headset project from Conix Research Center, Computing On Network Infrastructure for Pervasive Perception, Cognition and Action*, 2020. 2, 9

[46] R. K. Orosco, B. Lurie, T. Matsuzaki, E. K. Funk, V. Divi, F. C. Holsinger, S. Hong, F. Richter, N. Das, and M. Yip. Compensatory motion scaling for time-delayed robotic surgery. *Surgical endoscopy*, 35:2613–2618, 2021. 9

[47] L. Pantel and L. C. Wolf. On the impact of delay on real-time multiplayer games. In *Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video*, pp. 23–29, 2002. 1, 3

[48] K. S. Park and R. V. Kenyon. Effects of network characteristics on human performance in a collaborative virtual environment. In *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, pp. 104–111. IEEE, 1999. 3, 4, 9

[49] A. Pavlovych and C. Gutwin. Assessing target acquisition and tracking performance for complex moving targets in the presence of latency and jitter. In *Proceedings of Graphics Interface 2012*, pp. 109–116. 2012. 2

[50] M. Perez, S. Xu, S. Chauhan, A. Tanaka, K. Simpson, H. Abdul-Muhsin, and R. Smith. Impact of delay on telesurgical performance: study on the robotic simulator dv-trainer. *International journal of computer assisted radiology and surgery*, 11(4):581–587, 2016. 2

[51] J. Psotka. Immersive training systems: Virtual reality and education and training. *Instructional science*, 23(5):405–431, 1995. 1

[52] K. Raaen, R. Eg, and C. Griwodz. Can gamers detect cloud delay? In *2014 13th Annual Workshop on Network and Systems Support for Games*, pp. 1–3. IEEE, 2014. 1, 3, 9

[53] F. Richter, Y. Zhang, Y. Zhi, R. K. Orosco, and M. C. Yip. Augmented reality predictive displays to help mitigate the effects of delayed telesurgery. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 444–450. IEEE, 2019. 9

[54] G. Robertson, M. Czerwinski, and M. Van Dantzich. Immersion in desktop virtual reality. In *Proceedings of the 10th annual ACM symposium on User interface software and technology*, pp. 11–19, 1997. 2

[55] M. Rohde, L. C. Van Dam, and M. O. Ernst. Predictability is necessary for closed-loop visual feedback delay adaptation. *Journal of vision*, 14(3):4–4, 2014. 1, 3

[56] J. P. Rolland and H. Fuchs. Optical versus video see-through head-mounted displays in medical visualization. *Presence*, 9(3):287–309, 2000. 1

[57] C. Schaefer, T. Enderes, H. Ritter, and M. Zitterbart. Subjective quality assessment for multiplayer real-time games. In *Proceedings of the 1st Workshop on Network and System Support for Games*, pp. 74–78, 2002. 3

[58] J. Schulze, A. Forsberg, A. Kleppe, R. C. Zeleznik, and D. H. Laidlaw. Characterizing the effect of level of immersion on a 3d marking task. In *proceedings of HCI International*, vol. 5, 2005. 2

[59] R. H. So and G. K. Chung. Sensory motor responses in virtual environments: Studying the effects of image latencies for target-directed hand movement. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 5006–5008. IEEE, 2006. 2, 8

[60] A. Steed. A simple method for estimating the latency of interactive, real-time graphics simulations. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pp. 123–129, 2008. 1

[61] D. Ta Kim and D. Chow. The effect of latency on surgical performance and usability in a three-dimensional heads-up display visualization system for vitreoretinal surgery. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 260(2):471–476, 2022. 3, 4, 8

[62] D. S. Tan, D. Gergle, P. Scupelli, and R. Pausch. With similar visual angles, larger displays improve spatial performance. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 217–224, 2003. 2, 3

[63] J. T. Verhey, J. M. Haglin, E. M. Verhey, and D. E. Hartigan. Virtual, augmented, and mixed reality applications in orthopedic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 16(2):e2067, 2020. 2

[64] A. Vovk, F. Wild, W. Guest, and T. Kuula. Simulator sickness in augmented reality training using the microsoft hololens. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–9, 2018. 8

[65] T. Waltemate, F. Hülsmann, T. Pfeiffer, S. Kopp, and M. Botsch. Realizing a low-latency virtual reality environment for motor learning. In *Proceedings of the 21st ACM symposium on virtual reality software and technology*, pp. 139–147, 2015. 3

[66] T. Waltemate, I. Senna, F. Hülsmann, M. Rohde, S. Kopp, M. Ernst, and M. Botsch. The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*, pp. 27–35, 2016. 1, 2, 3, 9

[67] M. R. Wilson, J. M. Poolton, N. Malhotra, K. Ngo, E. Bright, and R. S. Masters. Development and validation of a surgical workload measure: the surgery task load index (surg-tlx). *World journal of surgery*, 35(9):1961–1969, 2011. 4

[68] S. Xu, M. Perez, K. Yang, C. Perrenot, J. Felblinger, and J. Hubert. Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dv-trainer® simulator. *Surgical endoscopy*, 28(9):2569–2576, 2014. 1, 2, 3, 4

[69] S. Xu, M. Perez, K. Yang, C. Perrenot, J. Felblinger, and J. Hubert. Effect of latency training on surgical performance in simulated robotic telesurgery procedures. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 11(3):290–295, 2015. 9

[70] K. Yarrow, I. Sverdrup-Stueland, W. Roseboom, and D. H. Arnold. Sensorimotor temporal recalibration within and across limbs. *Journal of Experimental Psychology: Human Perception and Performance*, 39(6):1678, 2013. 1

[71] F. Zheng, T. Whitted, A. Lastra, P. Lincoln, A. State, A. Maimone, and H. Fuchs. Minimizing latency for augmented reality displays: Frames considered harmful. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 195–200. IEEE, 2014. 9