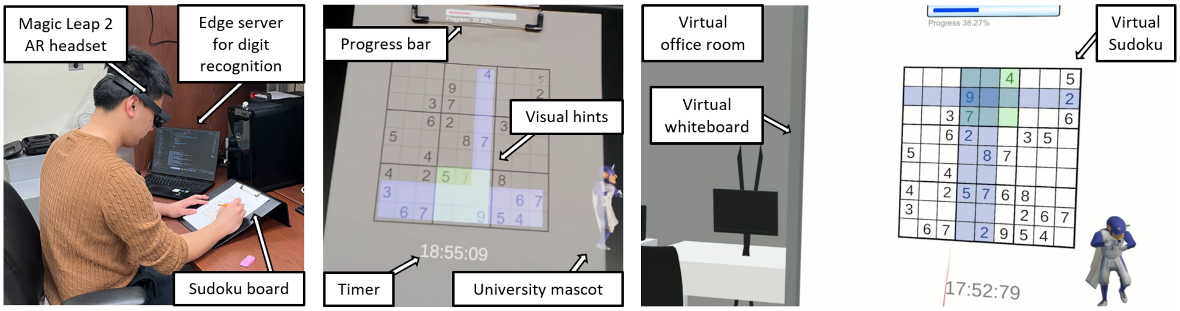
"Looking" into Attention Patterns in Extended Reality: An Eye Tracking–Based Study

Zhehan Qu*
Department of Computer Science
Duke University

Ryleigh Byrne†
Department of Electrical and Computer Engineering
Duke University

Maria Gorlatova‡
Department of Electrical and Computer Engineering
Duke University



(a) AR study setup (b) AR Sudoku Helper app interface (c) VR simulation of the AR Sudoku Helper app

Figure 1: Overview of the Sudoku Helper apps we developed for the study on user attention patterns in AR and VR environments. (a) The AR study setup with a participant wearing the Magic Leap 2 AR headset. (b) The interface of the AR Sudoku Helper app, showing the visual hints overlaid on a Sudoku puzzle, the progress bar, the timer and the university mascot, Duke Blue Devil. (c) The VR simulation of the AR app, showing the virtual office setup with the Sudoku puzzle placed on a whiteboard.

Abstract

Virtual reality (VR) simulations have been adopted to provide controllable environments for running augmented reality (AR) experiments in diverse scenarios. However, insufficient research has explored the impact of AR applications on users, especially their attention patterns, and whether VR simulations accurately replicate these effects. In this work, we propose to analyze user attention patterns via eye tracking during XR usage. To represent applications that provide both helpful guidance and irrelevant information, we built a Sudoku Helper app that includes visual hints and potential distractions during the puzzle-solving period. We conducted two user studies with 19 different users each in AR and VR, in which we collected eye tracking data, conducted gaze-based analysis, and trained machine learning (ML) models to predict user attentional states and attention control ability. Our results show that the AR app had a statistically significant impact on enhancing attention by increasing the fixated proportion of time, while the VR app reduced fixated time and made the users less focused. Results indicate that there is a discrepancy between VR simulations and the AR experience. Our ML models achieve 99.3% and 96.3% accuracy in predicting user attention control ability in AR and VR, respectively. A noticeable performance drop when transferring models trained on one medium to the other further highlights the gap between the AR experience and the VR simulation of it.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 Introduction

The deployment of AR applications in diverse environments to provide real-time information or guidance [67, 63, 9, 15] is becoming a practical reality. While AR has shown its potential to enhance user experience and task performance, the integration of AR content with real-world tasks raises concerns about the potential negative impacts of AR content on user attention. Visual clutter or obstruction of critical information caused by virtual contents may lead to suboptimal attentional states, such as attention tunneling [12, 13, 61] or distraction [26]. Commercial AR applications can add complexity to intended user attention; informative virtual contents and advertisements' virtual contents may compete for attention. Compared to traditional displays such as smartphones and computer screens, virtual contents in AR applications can have more intense, long-lasting impacts on users' attention, that are difficult to avoid spatially and temporally [70]. Therefore, it is crucial to understand the task-detrimental effects of AR content on user attention and task performance.

Fortunately, eye tracking has been widely adopted in modern XR headsets [47], making it possible to "look" into user attention patterns. Attention, a cognitive process directing neurocognitive resources toward behavioral goals, involves multiple brain networks [11, 31]. Monitoring the dorsolateral prefrontal cortex (PFC) helps differentiate attentional states [20]. Eye tracking, a standard method for analyzing attention, reflects how the PFC generates visual spatial attention signals, which are then translated into retinocentric signals and conveyed to the frontal eye field [29]. Attention researchers have leveraged such link between eye movements and attention by analyzing the spatial allocation of gaze, metrics of fixations (maintenance of eye gaze on a single location), and saccades (instantaneous and ballistic changes of the eyes between fixation points), as well as many other gaze events. Prior research has shown that spatial attention drives gaze events such as time-and-space-instructed saccades [23] and longer fixations at a spatial location indicate higher processing efforts towards that focus of attention [21]. These findings motivate the usage of eye tracking data to evaluate attention patterns during AR-assisted tasks [43, 58, 60].

To enable AR quality of experience (QoE) evaluation at scale, VR simulations for AR [32, 35, 50] have been proposed to provide controllable and extendable test environments for running tests of AR apps under diverse situations. Although widely used to justify the effectiveness of AR applications, it remains unclear whether VR simulations replicate the same effects that AR apps have on users.

*e-mail: zhehan.qu@duke.edu
†e-mail: ryleighbyrne@gmail.com
‡e-mail: maria.gorlatova@duke.edu

While users were found to maintain close performance and self-reported similar experiences in VR simulations [32], their behaviors in certain scenarios can be quite different [7]. Given the limitations in hardware, interaction methods, and the challenges in replicating the physicality and real-world content in VR, the lack of objective user state measurements is particularly concerning when using VR simulations for AR QoE evaluations. This calls for investigation into the validity of evaluating user attention patterns in AR with VR simulations to understand the impact of XR on users.

In this work, we developed a Sudoku Helper app to investigate how virtual contents affect user attention via eye tracking, for users wearing AR and VR headsets and with different attention control abilities. The app provides real-time guidance for solving the puzzle and introduces potential distractors at various task stages, enabling the comparison of user attention patterns under different AR and VR conditions. We conducted IRB-approved user studies with two disjoint groups of 19 participants for the AR and VR apps, where we collected eye tracking data and user attention labels. We analyzed the eye tracking-based metrics including fixations, saccades, and region-of-interest (ROI)-based fixation allocation, and trained transformer-based ML models [73] on eye tracking data to predict the presence of the distractors and the users' ability to control their attention. Our results show that the AR app had a statistically significant impact on user attention by increasing the proportion of fixated time (PFT) while the VR app had the opposite effect of decreasing PFT, indicating a potential discrepancy between VR simulations and AR. Evaluations of the ML models not only showed the potential of using eye tracking data for user attention pattern recognition and predicting personal attention control ability, but also corroborated the gap we found between AR and VR simulations, calling for more design efforts before using VR simulations for quantitative QoE evaluation. The code is available on Github[1]. Our key contributions are summarized as follows:

- We created an app for Sudoku solving in both AR and VR environments that provides step-by-step guidance and potential distractions, considered engaging by 94.7% participants.
- We directly compared attention patterns captured via eye tracking in AR and VR and found that VR simulation can induce higher perceptual load and decrease user focus, while cognitive load increased in AR.
- We trained ML models on eye tracking data to predict the existence of distractors and user attention control ability. The model performance drop when transferred between AR and VR further highlights the gap between AR and VR simulation.

The rest of the paper is arranged as follows. In Section 2 we cover related work, then in Section 3 we describe the app we developed. We then present the user study design in Section 4, followed by the analysis of the collected data and machine learning performance on those data in Section 5. We discuss the limitations and future work in Section 6, and conclude the paper in Section 7.

2 Related Work

Impact of AR on user attention. AR can affect user attention in distinct ways compared to traditional information medium. Techniques like omnidirectional attention funnel [4, 53] use spatial cursors to rapidly direct user attention to tracked objects, and for head-mounted AR eye tracking can be further utilized to provide adaptive guidance [52]. On the other hand, the potential negative impacts of AR on user attention that come together with security risks brought by buggy or malicious applications [2, 34] have also been reported. AR guidance was found to induce safety risks for drivers [64], while in the healthcare domain attention tunneling [12, 13, 61] and distraction [26] caused by the AR guidance system have been found to be detrimental to task performance, though no justification was provided on the determination of those suboptimal attentional states. Our work additionally leverages eye tracking data to quantitatively analyze the attention patterns.

User context sensing with eye tracking. A rich body of work on user context sensing is based on eye tracking, including emotion recognition [59, 62, 74], engagement detection [11, 57] and biometric identification [24, 40, 41]. These works used gaze points and pupil sizes as input to train machine learning models either with hand-crafted features for traditional ML models or with raw gaze data for deep learning models such as convolutional neural networks (CNNs). Recently, the idea of user context sensing with eye tracking has been extended to AR applications, as CAPturAR [68] leveraged egocentric camera feeds to provide fine-grained user activity recognition, and GazeGraph [33] used gaze data to detect one of six predefined activities performed by the users. Visual attention has also been investigated in AR applications when advertisements are displayed [72] and in a simulated driving scenario [16], though those works only focused on the spatial distribution and first-order statistics of the gaze data. Our work further extends the use of eye tracking to the recognition of user attention patterns in AR applications, training an end-to-end transformer-based model on the collected eye tracking data to predict whether the user is being distracted and the ability of the user to control their attention.

VR simulations of AR. VR simulations of AR have been recognized as a useful technique for evaluating AR applications in a more flexible manner [32, 35, 50], as they can provide fully controlled conditions, bypass hardware limitations of current AR devices (e.g., small field of view (FOV) [51]) and guarantee safe access to diverse environments that are hard to access in the real world. However, despite the advanced rendering quality provided by modern game engines, even highly realistic virtual environments still lack the physicality and real-world context inherent to AR applications, not to mention the differences in hardware and interaction methods [35]. In the context of VR eye tracking, gaze behaviors in VR have been studied in comparison to those in the real world among a number of tasks. Despite slight differences in gaze spatial distribution, human visual behaviors in VR has been proved to align with those in reality [3, 14]. AR shares similar characteristics in FOV [51, 69] with VR, and both AR and VR applications can lead to decreased blink rate [27, 28, 42] that might end up causing eye fatigue, which can also increase blink rate after prolonged usage [17]. To the best of our knowledge, our work is the first to directly compare attention patterns revealed by eye tracking in AR and VR, and our results raise concerns about investigating user attention in VR simulations as an all-encompassing evaluation of the AR counterparts.

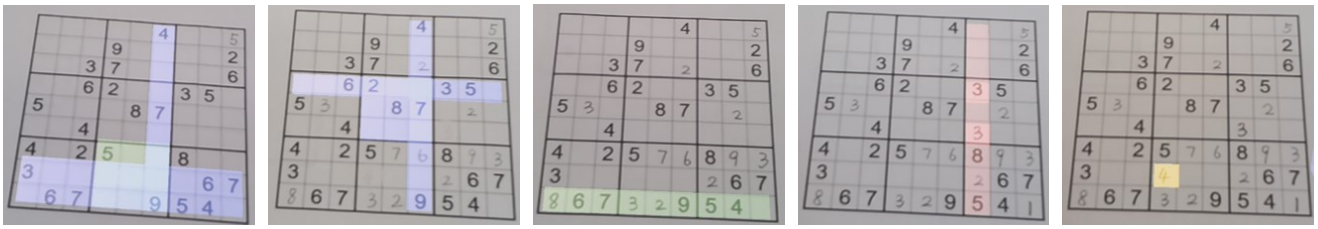
3 Task Formation for Attention Pattern Analysis in Extended Reality

To investigate user attention patterns when using XR headsets, we propose a task of *Sudoku solving* for attention pattern analysis, representing AR applications that provide guidance while being potentially distracting at the same time. Sudoku is a logic-based, combinatorial number-placement puzzle whose objective is to fill a 9×9 grid with digits so that each column, row, and 3×3 box contains all of the digits from 1 to 9. Solving a Sudoku is an attention-demanding task that leaves space for guidance. Additionally, Sudokus are classified into well-established difficulty levels, so that we can control the cognitive load implied by the task to avoid bias among trials. We develop our AR and VR apps on a Magic Leap 2 and an HP Reverb G2 Omnicept Edition, respectively, of which we will introduce the implementation details in the following sections.

3.1 AR Sudoku Helper Application

The workflow of our developed AR Sudoku Helper app is shown in Figure 3. The AR app is designed with an edge server [30, 38], running in geographic proximity of the AR headset, to provide necessary computational power for digit recognition and hint genera-

[1] https://github.com/Duke-I3T-Lab/XR_Attention_Sudoku



(a) Type 1: last remaining cell (b) Type 2: last possible digit (c) Type 3: last free cell (d) Type 4: warning w/ reference (e) Type 5: warning w/o reference

Figure 2: Examples of 5 types of hints to help solve the Sudoku. (a) A 7 should be filled in the last remaining cell in the bottom-middle box; (b) The last possible digit in the top-right of the center box is 1; (c) A 1 should be filled in the last free cell of the last row; (d) A warning with reference to another 3 in the 7th column; (e) Though not clear from the current puzzle, the 4 in yellow does not match the solution.

tion. At the beginning of each round of interaction, our app uses the RGB camera in the front of the AR headset to capture an image of the Sudoku puzzle and send it to the server for the recognition of the puzzle and the digits. The server then computes the current progress of the user and generates a hint for the next step, which will be sent back to the AR headset and rendered as virtual contents. The user will continue to add digits until the puzzle is completed. Throughout the procedure, eye tracking data, gaze-targeted AR contents and whether a distraction is present will be recorded. Next we will separately introduce the server (backend edge server) and the client (AR headset) components of the app.

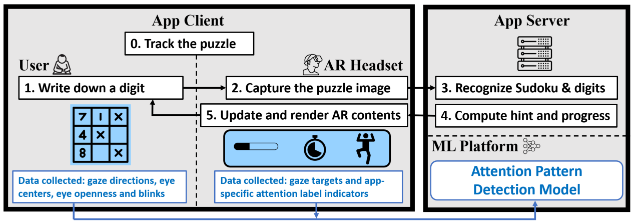

App Client
0. Track the puzzle
User
AR Headset
1. Write down a digit
2. Capture the puzzle image
5. Update and render AR contents
Data collected: gaze directions, eye centers, eye openness and blinks
Data collected: gaze targets and app-specific attention label indicators
App Server
3. Recognize Sudoku & digits
4. Compute hint and progress
ML Platform
Attention Pattern Detection Model

Figure 3: Workflow of the AR Sudoku Helper app.

3.1.1 Server Implementation

The edge server is in charge of generating hints that guide the users to complete the Sudoku given an image of the puzzle. The following three modules are involved in the server implementation, with the first two inspired by an online tutorial [55]:

Sudoku puzzle detection and grid extraction. Using the Python OpenCV library [5], we first extract the largest convex contour with 4 corners and apply a four-point perspective transform to obtain a top-down birds-eye view of the puzzle. We then divide the puzzle into 81 equal-sized cells to extract the digits in each cell.

Digit recognition. An Efficient-CapsNet [44] model is applied for digit recognition. If new digits are recognized, the progress (the percentage of the puzzle being filled) will be updated and the recognized puzzle will be sent to the hint generation module.

Hint generation. The hint generation module takes the recognized puzzle as input and generates hints correspondingly. If the recognized digit matches the solution of the puzzle, the server will generate the hint for the next valid move, in one of the three types [6]:

1. *Last remaining cell*. For one specific digit that is absent in one box, row or column, by eliminating cells that are in the same row, column or box of another instance of that digit, there is only one remaining cell that can be filled with that digit.
2. *Last possible digit*. For one of the empty cells, based on the existing digits in its row, column, and box, there is only one possible number that can be filled in that cell.
3. *Last free cell*. Only one cell is free in a row, column or box.

In cases where the recognized digit does not match the solution, the server will generate a negative hint, which highlights the cell that the user filled in incorrectly with (type 4) or without (type 5) a reference digit. Examples of all types of hints are shown in Figure 2. The server communicates with the client via TCP, through which the AR headset sends the raw captured image of the puzzle to the server, and the server responses with a message containing the current progress and the hint to be displayed.

3.1.2 Client Implementation

The client, namely the AR headset, is in charge of capturing the image of the Sudoku puzzle, rendering AR contents based on server responses and collecting eye tracking and app-specific data.

Sudoku image capturing. We use Vuforia image target detection to track the puzzle image, which remained reliable even when the user adds more digits, as the Sudoku image itself is full of edges and corners that made it feature-rich. Images of the puzzle are sent to the server every 1.5 seconds to provide a smooth experience for the user while ensuring timely recognition of the digits.

AR guidance and distractors. The client interface is shown in Figure 1b. Upon receiving the response from the server, the client will render the transparent *visual hints* in the form of rectangle covers on the puzzle, in the color of blue and green for positive hints and red or yellow for negative hints (see Figure 2 for examples). We additionally add the following AR components in the app:

- *A progress bar* on top of the puzzle board, showing the percentage of cells already correctly filled in the puzzle.
- *A timer* at the bottom of the puzzle board, counting down from a 20-minute limit.
- *A university mascot* (Duke Blue Devil) dancing at the bottom-right of the puzzle. When a mistake is recognized, the mascot will shake its head. Additionally as a distractor, every 1.5s the mascot has a 7% chance to start *running* around the puzzle.
- *Audio hints* in the form of a man's voice of the digit to fill being played repeatedly when a hint of type 2 or 3 is received.

The user interface closely resembles that of a commercial Sudoku helper app, featuring elements that convey information, enhance user enjoyment, and subtly reveal the app's creator. We envision this design as representative of a typical AR application, which offers guidance while maintaining an element of distraction.

Data collection. The headset records eye tracking data, including gaze directions, eye center positions and, eyes open amount at 60Hz. For the identification of different periods and attention labels during the session, whether the audio hint is activated and whether the university mascot is running will also be recorded at each timestamp. We also record the gaze targets of the user, including the *puzzle* board, the *progress bar*, the *timer*, the *mascot*, and the *hints* overlaid on the puzzle board.

3.2 VR Simulation of the AR Sudoku Helper Application

To explore the validity of evaluating attention patterns in AR with VR simulations, we developed a VR Sudoku Helper app that mimics the AR app. While sharing the same task and virtual elements as described in Section 3.1.2, due to the nature of the fully virtual world and hardware limitations of the headsets we created the app such that it is different from the AR app in the following aspects:

Interaction method. Digit input in the VR app relies on the controller. Users can point the ray from the controller at a cell and use the trigger at the back to select. The joystick is used to scroll between digits, and the trigger is then used to confirm.

Puzzle size and placement. To accommodate the controller-based interaction, a larger puzzle in VR is vertically positioned on a virtual whiteboard in front of the user, perceived to be slightly beyond an arm's length with a diagonal viewing angle of 35°, intentionally placed far away to avoid depth perception issues due to different user poses. The puzzle horizontally placed in AR has a diagonal viewing angle of 15°–25° depending on the user's writing pose.

Hint generation. In VR the puzzle and every entered digit are readily known to the server, making the hint generation time shorter and more stable than that in AR without the need to run the digit recognition model.

Data collection. The AR headset records eye tracking data at 60Hz. The VR headset is equipped with Tobii eye tracking sensors that collect gaze directions and pupil dilation at 120Hz, but do not record eye center positions or eye open amounts. Both devices' eye tracking accuracy is within 1 degree.

While we acknowledge that the differences between the AR and VR apps can introduce confounding factors to the comparison of user attention patterns, we note that, unfortunately, current VR simulations cannot fully represent AR experiences (e.g., controller must be used for high-precision input [25] in VR. Scene understanding functions, such as object detection, in AR often lead to unstable response times [18]), and it is difficult for VR simulations to replicate this instability. We believe that the comparison of AR and VR user attention can still provide valuable insights under the current circumstances, but further research is needed to ground the findings and further inform the design of VR simulations.

3.3 System Implementation Details

We developed both apps in Unity 2022.3.6f1. The server was implemented with Python 3.8.11 and the Efficient-CapsNet model was implemented with TensorFlow [1]. We adopt the same Sudoku puzzles from puzzles.ca [49] for both studies. The average latency of hint visualization, measured from the time the image is captured to the point when hint gets rendered on the client, is 254ms in AR and 11ms in VR. As the average time for a user to fill in a cell was typically greater than 2s, the latency was considered acceptable for hint generation in AR. Subjective feedback from the users on the latency can be found in Section 5.1.

4 AR and VR Attention Analysis User Studies

In order to evaluate the impacts of AR on user attention and the validity of evaluating them in VR simulations, we conducted two user studies approved by the Duke University Institutional Review Board. The objective of the studies is to gather eye tracking data and user attention labels from participants as they solve Sudoku puzzles in both AR and VR environments. We will then analyze these data to determine whether the two media have varying impacts on users with different levels of attention control ability. Machine learning models were also trained on the data for the prediction of the presence of distraction and personal attention control ability.

4.1 Study Procedure

The same procedure was employed in both studies. Upon arrival, participants were first asked to read and sign the consent form. Next, participants were instructed to complete the attention control test known as the Flanker Squared task proposed by Burgoyne et al. [8], in which they would consecutively identify the central target stimulus amidst surrounding distractors in a 90s period to obtain a score. Participants would then fill in a questionnaire about their familiarity with XR and Sudoku and be introduced to the task. A demonstration video would then be played while the researcher further helped with interpreting the guidance and distractors. After eye calibration, participants would first work on an easy-level puzzle for 3 minutes wearing the headset. This trial was unguided, meaning that AR users would only see the Sudoku printed on a physical paper, and VR users would see the virtual Sudoku alone. Then they would perform a 20-minute session of solving a hard-level puzzle with all virtual contents present. Users were instructed to try not to make mistakes during the task to ensure that attention was paid properly. The session would end immediately after the puzzle was solved or the 20-minute limit was reached. Post-task, participants filled out a questionnaire and received a Duke souvenir as compensation. Each session lasted approximately 45 minutes.

4.2 Participants and Environment

We recruited 19 different participants from our university campus for each study via email, electronic flyer and poster. Users would indicate their preferences when signing up and we balanced the two groups if they did not express a preference. Among the 19 AR study participants (mean age: 22.8 ± 3.3 years, range: 18–32 years; 6 female), 7 wore glasses, 2 reported to have used AR headsets at least once or twice a week before, 8 had worn an AR headset once or twice, and 9 had never worn an AR headset. 14 of them self-reported to be at least "moderately familiar" with Sudoku solving, while 10 had solved Sudoku puzzles with hints. For the 19 VR study participants (mean age: 24.7 ± 5.3 years, range: 21–42 years; 8 female), 8 wore glasses, 3 had used VR headsets at least once or twice a week before, 8 had worn a VR headset once or twice, and 8 had never worn a VR headset. 13 were at least "moderately familiar" with Sudoku solving, while 12 had solved Sudoku puzzles with hints. The AR study was conducted in a quiet room with controlled lighting conditions, while the VR study was conducted in a different quiet room without lighting being controlled.

5 Analysis and Results

We collected data from 38 participants and present the results in this section. The average completion time was 17.4 ± 2.8 mins for the AR study and 15.4 ± 3.4 mins for the VR study. The average attention control score (ACS, higher indicates better attention control ability) for the AR group was 39.0 ± 10.8, while the VR group scored 36.2 ± 8.8. In the following sections we use non-parametric tests for the analysis of the data (Mann-Whitney U tests for comparisons between AR and VR and Wilcoxon signed-rank tests for paired comparisons within each group) due to the small population size of our data.

5.1 Questionnaire Results

In the post-study questionnaire we asked the participants whether they found the experience to be engaging (questions adopted from the Game Experience Questionnaire [48]), whether they found each of the virtual contents to be useful or distracting, whether they made mistakes during the session and reasons for the mistakes. The answers to those questions are summarized in Figure 4. Out of the five-point Likert scale answers, we group "agree" and "strongly agree" as positive responses and report the *positivity rate* of those responses in the following sections. The participants' free-text response are quoted with the participant number, *P*.

Engagement. Both AR and VR users reported high engagement level with the task, with a positivity rate of 94.7% on "being fully occupied" for both studies and only 10.5% and 5.3% for "thought about other things" in the AR and VR study, respectively.

Usefulness of XR contents. The AR users reported positivity rates of 78.9% for the usefulness of visual hints, 21.1% for the progress bar, 31.6% for the timer, 5.2% for the mascot's pose change and 89.5% for the audio hints. In VR the positivity rate was 68.4% for the visual hints, 21.1% for the progress bar, 63.2% for the timer, 10.5% for the mascot's pose change and 68.4% for the audio hints. We found the contents that are not directly related to the puzzle solving task to be reported less useful than the visual and audio hints. The mascot was reported to be the least useful in both studies, which is consistent with our design of the mascot more as a distractor than a guidance. The higher positivity rate on the usefulness of the timer and progress bar indicated that the users might

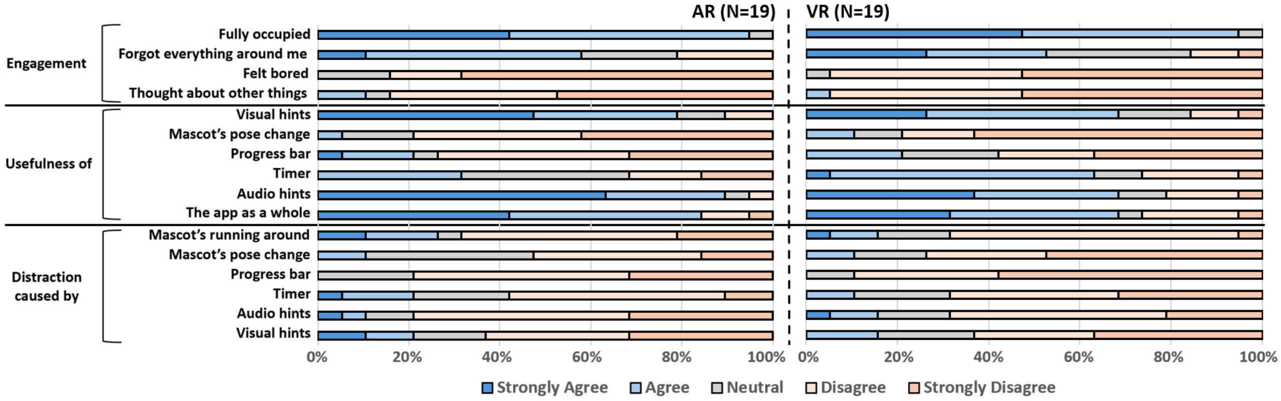


Figure 4: Survey responses indicating user engagement levels and perceived usefulness or distraction of each XR content element.

still appreciate the additional information provided by XR that is not directly related to the task. Overall, 84.2% and 68.4% users were positive towards the Sudoku Helper app being useful in AR and VR, respectively, though the difference between their ratings was not found significant (p = .36) through a Mann-Whitney U test.

Distraction caused by XR contents. The proportion of AR users that agreed or strongly agreed with each element being distracting was 21.1% for the visual hints, 0% for the progress bar, 21.1% for the timer, 10.5% for the mascot's pose change, 26.3% for the mascot's running around the puzzle and 10.5% for the audio hints. For VR, we found the proportions to be 15.8% for the visual hints, 0% for the progress bar, 10.5% for the timer, 10.5% for the mascot's pose change, 15.8% for the mascot's running around the puzzle and 15.8% for the audio hints. We found the mascot's running behavior that was designed as a distractor, was reported to be overall not distracting, most likely due to the fact that users were highly engaged in the Sudoku solving task, being in the state of flow [65] that made them ignore distractions. For the 5 AR users that agreed or strongly agreed to the mascot's running being distracting, 4 of them had an ACS $<$ 39 (mean ACS of the AR group); while all 3 VR users that found it distracting had an ACS $<$ 36.2 (mean ACS of the VR group), showing that the users with lower attention control ability were more likely to be affected by the distractors. However, although self-reported results indicated almost no effect from the distractors, quantitative analysis of the eye tracking data revealed different attention patterns during those distracted periods, which we will discuss in Section 5.2.4.

Subjective feedback on hint generation. No VR users mentioned anything about the hint generation in their free-text responses. Two AR users mentioned that the hints were "slow," with P7 saying that "*The only thing was sometimes it took some time for the headset to recognize the number I wrote*" and P13 mentioning that it was "*Too slow for someone who already knows how to play*." At the same time, P19 mentioned that "*It was easy to use,*" while P6 said that "*I liked the hint they were very timely and accurately overlayed.*" The mixed feedback on the hint generation speed did show a difference in the user experience between the AR and VR studies, which we will further discuss in Section 6.

Mistakes. 12 and 18 participants self-reported to have made mistakes during the session for the AR and VR studies, respectively. However, only 2 users in each study considered "not paying attention" as the reason for their mistakes, while 9 and 10 users attributed their mistakes to "*being confused by the hints*" in AR and VR, respectively. Given that the users did not go through a guided tutorial trial where they can familiarize themselves with the hints prior to the 20-minute session, the level of understanding of the hints might be a factor that affected their performance. We envision fine-grained analysis on periods before mistakes (which can be indicators of more subtle suboptimal attentional states) to be enabled by creating a more controllable task where suboptimal attentional states would be the only reason for mistakes.

5.2 Attention Pattern Analysis with Gaze Data

5.2.1 Setup for Gaze Event Extraction

Prior to the extraction of gaze events such as fixations, saccades and smooth pursuits, the VR data were first downsampled from 120 Hz to 60Hz to match the AR data frequency. We used the I-VT algorithm [56] with a velocity threshold of 30 deg/s to detect fixations. Gaze movements above the threshold were considered as saccades, except for smooth pursuits which we defined as gaze movements that did not target both on-puzzle and off-puzzle targets with a velocity between 30 deg/s and 100 deg/s (thresholds were chosen following Tobii's practice [46]). We reported fixation-related metrics after aggregating smooth pursuits with fixations, as they both indicate the focus of visual attention [10]. For ROI analysis, we defined five ROIs as *hints*, *progress bar*, *timer*, *mascot* and *puzzle*, and analyzed fixations on each ROI.

In other to compare differences in user attention patterns during their XR experiences, we extracted five representative periods out of each user study session as described below:

- The ***un**guided period (UG)* when the user was solving the easy-level puzzle without any other virtual contents;
- The entire session of solving the hard-level puzzle ***u**sing* the Sudoku Helper ***a**pp (UA)*;
- The ***n**o-**d**istraction period (ND)* during which the mascot was not running and the audio hint was not played;
- The ***m**ascot-**r**unning period (MR)* when the mascot was running around the puzzle;
- The ***a**udio-**p**laying period (AP)* when the audio was played.

We analyzed metrics related to fixations and saccades (Figure 5) and ROI-based fixation distributions during these periods. We conducted within-group comparisons between UG and all other periods to test for the effect of virtual elements in the app on gaze patterns, plus ND vs. MR and ND vs. AP for the effect of specific distracting stimulus (total 6), as well as between AR and VR users in each period (total 5) to test for the effect of media.

5.2.2 Fixation Metrics

We computed mean fixation duration (MFD), fixation rate (FR) and the proportion of fixated time (PFT) for each period in AR and VR. Mixed-design ANOVA (with XR medium as the between-subject factor and period as the within-subject factor) showed a significant interaction effect [$F(4,144) = 47.06$, $p < .001$, $n_p^2 = .57$] and main effect of period [$F(4,144) = 11.23$, $p < .001$, $n_p^2 = .24$] on MFD, a significant interaction effect [$F(4,144) = 37.96$, $p < .001$, $n_p^2 = .51$] and main effect of period [$F(4,144) = 79.01$, $p < .001$, $n_p^2 = .69$] on FR, and a significant interaction effect [$F(4,144) = 27.72$, $p < .001$, $n_p^2 = .43$] and main effect of period [$F(4,144) = 59.58$, $p < .001$, $n_p^2 = .62$] on PFT, suggesting that the effects of XR contents on fixation were significant, and such effects

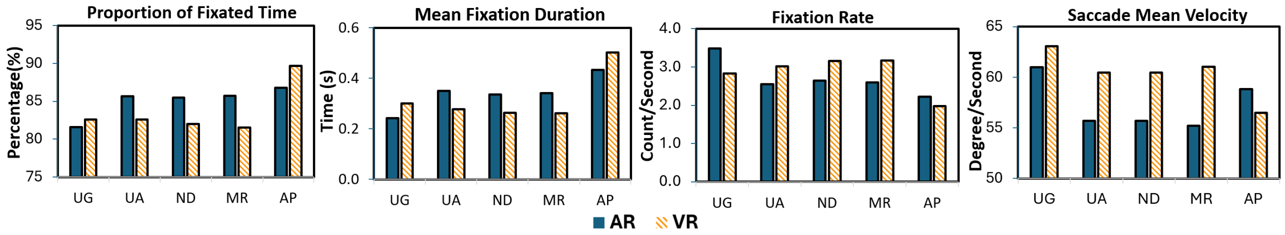


Figure 5: Metrics of gaze events in different periods. UG: unguided; UA: using app; ND: no distraction; MR: mascot running; AP: audio playing.

varied across XR medium. The effect of XR medium was not found significant on any of these metrics, but given the significant interaction effects, we also conducted between-group comparisons for each period [22]. Using Bonferroni correction, we set the significance level at $p < .0083$ for within-group comparisons and $p < .01$ for between-group comparisons (corrected from $p < .05$).

As shown in Table 1, distinct effects of virtual contents were found on MFD and FR of AR and VR users. In AR, *MFD* was found to *increase* in UA, while *FR decreased*. The average *proportion of fixated time (PFT)* also *went up* from 81.6% in UG to 85.7% in UA, with similar increases in ND (85.5%), MR (85.7%) and AP (86.8%). On the contrary, VR users had a *lower MFD* in UA except for the AP period, with an *increase in FR* except for the AP period as well, with the significant changes of FR from UA to MR found to be of opposite directions in AR vs. VR. When using the VR app, PFT did not change significantly except for the AP period, where a significant increase was observed. Based on research by Liu et al. [37], prolonged fixation duration and fewer fixation counts correlate with increased cognitive load (mental effort to complete a task). Conversely, shorter fixation duration and more fixation counts are associated with increased perceptual load (sensory demands on the perceptual system). Our findings indicate that the visual components of the AR and VR app had distinct effects: the AR app may raise cognitive load, while the VR app may heighten perceptual load.

Additionally, the *MFD* of the UG period was found to be significantly *longer* in VR than in AR ($p = .002$, $A = .21$) together with a *lower FR* ($p < .001$, $A = .81$), which aligns with prior findings that fixation durations are typically longer in VR [3] due to lack of objects to fixate at. This trend was reversed when the app was used and more visual contents were added, as *MFD* became *longer* in AR for the UA ($p < .001$, $A = .83$), ND ($p < .001$, $A = .87$) and MR ($p < .001$, $A = .85$) periods, with *lower FRs* and *higher PFTs* as well (all tested significant). Resutls indicate that AR experience largely differs from VR, and assuming VR simulation to always be a valid representation of AR experience might not be appropriate.

Furthermore, the effect of audio on fixation metrics was significant in both AR and VR, leading to an increase in MFD (AR: $p < .001$, $A = .33$; VR: $p < .001$, $A = .03$) and decrease in FR (AR: $p < .001$, $A = .70$; VR: $p < .001$, $A = .94$) when compared with the ND period, resulting in a significantly *higher PFT* in VR ($p < .001$, $A = .02$; AR had $p = .02$). Such results indicate that the audio hints had a medium-agnostic, attention-enhancing effect on the users, potentially increasing the cognitive load by urging the users to fill in the digit faster. In the future we plan to explore the impact of audio stimulus in a broader sense with more types of audio hints applied.

5.2.3 Saccade Metrics

The saccade mean velocity (SMV) was computed for each period. Mixed-design ANOVA showed a significant interaction effect [$F(4,144) = 9.00$, $p < .001$, $n_p^2 = .20$] and main effect of period [$F(4,144) = 10.91$, $p < .001$, $n_p^2 = .23$] on SMV, but the effect of XR medium was not found to be significant. The same Bonferroni correction was applied on SMV comparisons. We did not conduct pairwise comparisons between AR and VR saccade metrics due to differences in puzzle board sizes.

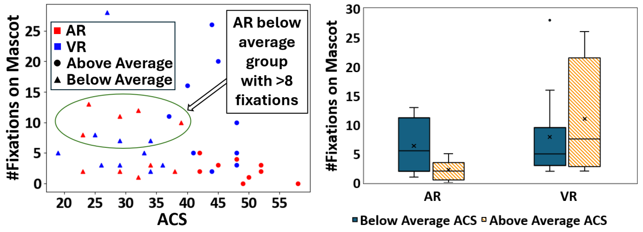
Table 1: Differences in fixation metrics between UG and other periods. Vargha and Delaney's A effect size is reported together with the p-values of the Wilcoxon signed-rank test. The effect size is interpreted as small (.56), medium (.64) and large (.71) if a decrease in the metric was found ($A > .5$), otherwise ($A < .5$) as large (.29), medium (.36) and small (.44) if an increase was found. Significant increases are marked in bold and decreases are marked in italics.

		Period							
XR	Metric	UA		ND		MR		AP	
		A	*p*	*A*	*p*	*A*	*p*	*A*	*p*
AR	MFD	**.11**	< .001	**.12**	< .001	**.11**	< .001	**.06**	< .001
	FR	*.89*	< .001	*.88*	< .001	*.89*	< .001	*.94*	< .001
	PFT	**.17**	< .001	**.18**	< .001	**.17**	< .001	**.16**	< .001
VR	MFD	.62	.28	.73	.03	.70	.02	**.13**	< .001
	FR	.37	.23	.27	.01	**.30**	.007	*.85*	< .001
	PFT	.50	.99	.58	.52	.59	.31	**.06**	< .001

We found SMV to decrease comparing the UG period with the UA period in both AR and VR, with an 8.7% decrease from the average 61.0 to 55.7 deg/s in AR ($p < .001$, $A = .80$) and a 4.1% decrease from 63.1 to 60.5 deg/s in VR ($p < .001$, $A = .63$), suggesting that more contents displayed in a constrained region tend to suppress fast saccades. Meanwhile, though not found significant in the current study, audio had opposite effects on SMV in AR and VR comparing ND to AP, as in AR audio made saccades faster from 55.7 to 58.8 deg/s ($p = .03$, $A = .37$) while in VR slower from 60.5 to 56.5 deg/s ($p = .01$, $A = .66$). We suspect that such difference might be due to the difference in how the audio hints in the real world and in VR were perceived, as in AR users may have the intuition to search for the source of the sound in the real world.

5.2.4 Fixations on ROIs

Starting from this section, we use "PFT" as the proportion of fixated time on each ROI to the *total fixation time of that user*. Mixed-design ANOVA shown that in the four periods where all ROIs were present, periods were found to have a significant effect on the PFT on the mascot [$F(3,108) = 26.13$, $p < .001$, $n_p^2 = .42$], and the interaction effect was also significant [$F(3,108) = 13.36$, $p < .001$, $n_p^2 = .27$]. We made within-group comparisons of ND vs. MR and ND vs. AP, as well as between AR and VR users in each period. Bonferroni correction was applied such that the significance level was set at $p < .025$ for within-group comparisons and $p < .0125$ for between-group comparisons. For the AR users, the PFT spent on the mascot was similar in all five periods (0.4% in UA and ND, 0.5% in MR and 0.3% in AP), while for the VR users there was a larger difference when the mascot was running (0.4% in UA, 0.1% in ND, 1.5% in MR and 0.2% in AP, ND vs MR had $p < .001$, $A = .001$). Note that the percentage reported is about the "fixated time" on the mascot, excluding short periods of gaze targetting the mascot but not fixating on it, and 1.5% indicated approximately 1s of fixated time on the mascot out of a 1min period, in a highly focused task like Sudoku solving. The effect of MR was found to be strong, and the difference between PFT on the mascot of VR and



(a) (b)

Figure 6: Scatterplot (a) and boxplot (b) of #fixations on the mascot during the MR period for AR and VR users with different attention control abilities. (a) >8 fixations on the mascot in AR were only observed among users with below-average ACS (green ellipse). (b) AR users with above-average ACS fixated less on the mascot than those with below-average ACS (not found significant in VR).

AR users in the MR period was significant (VR > AR, $p < .001$, $A = .83$), indicating a larger distraction effect in VR was caused by the mascot's running.

We additionally investigated how attention control ability would affect the users' distractibility, measured by the number of fixations they had on the mascot. As shown in Figure 6, in the MR period we found the AR users who scored below the average ACS of the AR group fixated 6.4 times on average on the mascot, while those above average only fixated 2.2 times ($p = .01$, $A = .27$), with all users who fixated on the mascot more than 8 times belonging to the below-average subgroup. However for the VR users, no significant difference was found in the number of fixations on the mascot between the two groups. Such finding agreed with the fact that in AR, the more cognitively demanding task made the mascot less likely noticeable, serving as a shield to distraction [19, 54], and that individual attention control ability led to different attention patterns in AR but not in VR.

5.3 Machine Learning on Eye Tracking Data

Using the eye tracking data collected, we trained ML models for two purposes: (1) predicting whether the user was going through a distraction brought by the XR content, namely if the mascot was running (the *MR* task), where we obtained the labels from our recorded data; and (2) predicting the user's attention control ability, where we categorized the users into three groups of high, medium and low attention control ability based on their scores in the Flanker Squared task (the *ACS* task). The thresholds for the three groups were chosen based on the reported mean score of 27.38 with a standard deviation of 13.96 in the study where the task was proposed [8]: we chose 41 (one standard deviation above the mean) and 27 (the mean) as the thresholds so that the number of users in each group was relatively balanced. We envision such tasks to be valuable for future XR applications for two reasons: first, knowing whether the user is distracted can help build attentive interfaces that can adapt to the user's attentional states; second, similar to previous work on attention-deficit/hyperactivity disorder (ADHD) prediction using gaze data [45], predicting the user's attention control ability can help the app to be personalized for different users, to avoid negative QoE during AR or VR usage.

Dataset. We extracted windows of 120 frames (≈ 2s) from the raw gaze data to formulate our time-series dataset for ML model training. Each data sample $\mathbf{X} \in \mathbb{R}^{w \times m}$ in the dataset is a sequence of $w = 120$ frames in which each frame contains m features. We performed tests on multiple combinations of available features. For the AR data, the full feature set includes gaze directions, eye centers, eyes open amount and gaze targets, while for VR we used gaze directions, pupil dilation and gaze targets. We grouped all visual AR contents into two categories: *useful*, including all hints overlaid on the puzzle board and *relevant*, for the progress bar, the timer and the university mascot that are part of the app but not directly helpful to the puzzle-solving task. Such grouping technique can be generalized to other AR applications that provide highly informative guidance while displaying less useful but still relevant contents. The "gaze target" feature is thus a one-hot variable about whether *useful* or *relevant* AR contents, or the *puzzle* was being targeted.

For each AR participant, eyes open amount was first normalized based on the data collected during the 3-minute easy puzzle solving period. Additionally, a large-scale VR eye tracking dataset named GazeBaseVR [39] was used for the evaluation of the effectiveness of pretraining. The dataset is composed of gaze directions and 3D eye centers collected at 250Hz. The entire dataset was first downsampled by 4 times, and split into windows of 2 seconds for pretraining. Since the pretraining dataset was collected in VR where the users primarily looked forward, we converted the collected AR gaze directions (where the users primarily looked downward at the puzzle) to the puzzle board's coordinate system prior to training. Such conversion also made our AR and VR datasets more comparable when transferring models from one to the other.

Model architecture. The MVTS-Transformer model [73] was used for the tasks due to its compatibility with time-series data and its ability to capture long-range dependencies. It is based on the transformer [66] architecture where all the features at each timestamp are first linearly projected to a higher-dimension vector and then fed into multiple transformer blocks to capture the temporal dependencies. The framework also enables unsupervised pretraining by mask-recovery, where features at each timestamp are randomly masked out and the model is trained to predict the masked features based on the rest of the features. In our implementation, only features that exist in both GazeBaseVR and the Sudoku dataset can be masked out during the pretraining stage, depending on which features were selected in the AR or VR data.

Model training and evaluation setup. Following the common practice for time-series data [36, 73], the data splitting process began by dividing each user's longitudinal data 8:1:1 into three time chunks, designated as training, validation, and test sets. From these segments, 2s windows were extracted and the resulting subsets from all users were then aggregated into three comprehensive datasets for training, validation, and testing. For all the tasks we performed, models were trained with the Adam optimizer using a learning rate of 0.001 and a batch size of 32 for 40 epochs, and we chose the best-performing model on the validation set for evaluation. Performance metrics including accuracy and macro-averaged one-vs-one AUC were reported on the the test set. Pretraining was performed on GazeBaseVR for 5 epochs using the same hyperparameters. We also performed ablation studies on different sets of features to evaluate the importance of each feature for the tasks.

Results. We trained models with different sets of features, with and without pretraining and report the results in Table 2. Note that due to the difference in the durations of each period and the specific population we had on attention control ability, the label distributions of these tasks were imbalanced. For the AR users, the non-MR period accounts for 71.8% of the data in the MR task, while for the ACS task 45.1% were classified "high"; for the VR users, the non-MR period accounts for 71.9% of the data in the MR task, while for the ACS task 49.5% were classified "high."

We found the model for predicting ACS to perform better, with the model trained with gaze directions, eye centers, eyes open amount and gaze targets achieving a 99.3% accuracy and 99.9% AUC on the AR data when fine-tuned on the pretrained model, showing the effectiveness of pretraining on large datasets. When training from scratch, the model trained without eyes open amount performed the best with a 98.8% accuracy and 99.9% AUC, indicating that all these features were beneficial for predicting user-level traits, showing great potential for using eye tracking data for user context inference in XR applications. The VR model also performed comparably well using gaze directions, pupil dilation and gaze targets, achieving a 96.3% accuracy and 99.9% AUC on the

ACS task. However for the MR task, the models performed only slightly better than guessing it to be the majority class, with AUC scores ranging around 52% to 58% across all models. We believe the model performance suffered from the fact that rarely did users have distinct responses to the mascot running behavior, leading to different labels assigned to similar gaze patterns. More powerful distractors that can lead to distinct gaze behaviors might be more effective, as we plan to implement in the future.

Table 2: ML model performance on predicting MR and ACS.

XR	Features	MR				ACS			
		Scratch		Fine-tune		Scratch		Fine-tune	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
AR	Dr+A+T	.721	.533	.733	.549	.953	.995	.945	.993
	Dr+C+T	.738	**.575**	**.757**	.561	.988	**.999**	.988	**.999**
	Dr+C+A+T	.725	.554	.729	.557	.984	**.999**	**.993**	**.999**
VR	Dr+Di+T	**.772**	.525	.764	**.565**	.950	.992	**.963**	**.996**

[1] Feature abbreviations: Dr: gaze directions; Di: pupil dilation; A: eyes open amount; C: eye centers; T: gaze targets.

[2] "Fine-tune" indicates fine-tuning after pretraining on GazeBaseVR.

One of the perspectives of VR simulations is that they can be used to collect large amounts of data for training models that can be later applied to AR applications (known as domain adaptation). We additionally evaluated the feasibility of transferring models between AR and VR (see Table 3) on the ACS task where the models on its own domain performed well. To make models transferable, they were trained only on shared features in AR and VR, namely on gaze directions and gaze targets (Dr + T) exclusively. We found a performance drop when transferring the models in both directions and for both tasks, in which the adaptation of the model trained from scratch on the VR data to AR suffered the most, whose accuracy dropped to 38.8% (-52.0% from its VR performance and -56.6% from the AR model) and AUC to 46.9% (-49.8% from its VR performance and -52.2% from the AR model). Such results indicate that the user attention patterns in AR and VR were distinct, and the models trained on one domain cannot be directly applied to the other domain. Another valuable finding is that the model that was pretrained on GazeBaseVR can retain better performance when transferred from AR to VR (the pretrained model was able to keep 50.2% accuracy vs. 30.1% for the model trained from scratch), suggesting that pretraining on a large-scale VR dataset can be beneficial for the model to learn generalizable features. If the tasks were more similar or the features were more aligned between GazeBaseVR and our AR/VR datasets, we believe that pretraining would help the model to retain acceptable performance when transferred between different domains, as a promising mitigation of domain gaps between AR and VR.

6 Discussion And Future Work

The results of our study indicate that our AR app had a prominent effect on enhancing user attention, which might not be replicated by the VR simulation. While the effect of increasing cognitive load (increasing MFD and decreasing FR) of our app might be specific to our task of Sudoku solving, the drastic difference in user gaze behaviors between AR and VR indicated that *the two media are not interchangeable in terms of accurately replicating user attention patterns*. We also found that the audio hints had a consistent effect on fixation metrics. However, the opposite effects on saccade metrics in AR and VR call for further investigation, likely with a study specifically tailored for audio stimuli in AR. We also found distractions to be less distracting than expected, and those in the form of audio hints might have even promoted user attention instead. Additionally, we found that ML models worked better for predicting user attention control ability than for predicting user attentional states, which can potentially be improved by using more powerful distractors that lead to more distinct gaze behaviors, for example by violating known good practices for attentive interfaces [47], such as violation of common fate and color manipulation [71] to force separation of attention.

Table 3: Model performance on the ACS task when transferred between AR and VR. Models were trained using shared features.

XR	Scratch		Fine-tune	
	ACC	AUC	ACC	AUC
AR	.8955	.9809	.9131	.9871
VR → AR	*.3883*	*.4693*	*.3586*	*.4750*
VR	.8088	.9352	.8717	.9639
AR → VR	*.3005*	*.4093*	*.5019*	*.5520*

While our findings on attention in XR are intriguing, there are still limitations that need to be addressed in the future. First, our task of Sudoku solving was designed to be generalizable, yet as a task that involves more mental than physical effort, the conclusions drawn from this study might not generalize to use cases that involve high mobility or high-precision targeting. Second, the participants we recruited were mostly from our university, which might not be representative of the general population, especially in terms of their attention control ability. Third, the VR app we developed did not replicate the AR app at "pixel-to-pixel" level, and had differences in interaction methods, puzzle size, hint generation latency and hardware specs. While in essence a precise replication of AR experiences in VR simulations remains elusive, we recognize that these disparities may have impacted user experience and, consequently, experimental outcomes (e.g., larger puzzles and heavier headset may lead to eye fatigue; digit recognition in AR can introduce instability). Further studies on VR simulations need to be carried out to understand how to successfully replicate both interaction and unpredictable scene understanding functions in AR. Finally, our ML models were trained and tested on a shuffled dataset with all users' data combined, yet for deploying such models to produce personalized contents or mitigate attention-detrimental AR experience, models with zero-shot inference ability on unseen users are desired. We will explore ways to generalize the model, not only to unseen users but also to other tasks, and eventually develop a plug-and-play model that can be applied to any custom-designed app for detecting and mitigating suboptimal user attentional states.

7 Conclusion

In this paper, we characterized user attention patterns in both AR and VR using eye tracking data with a custom-designed Sudoku Helper app. We found that AR contents had an effect of increasing cognitive load and enhancing attention, while opposite findings on gaze metrics in VR suggested that the two media are not interchangeable in terms of replicating user attention patterns. Though working well on their own domains, ML models trained for predicting user attention control ability were also found to not work properly when transferred between AR and VR domains, suggesting making inference on user attention patterns in AR based on VR simulations might not be appropriate. Our findings reveal intriguing differences between VR simulations and AR and serve as a starting point for the development of countermeasures for suboptimal attentional states in XR applications.

Acknowledgments

The authors want to thank Professor David Carlson for his valuable feedback on the work. This work was supported in part by NSF grants CNS-1908051, CNS-2112562 and IIS-2231975, NSF CAREER Award IIS-2046072, a CISCO Research Award, a Meta Research Award, and Defense Advanced Research Projects Agency Young Faculty Award HR0011-24-1-0001. This paper has been approved for public release; distribution is unlimited. The contents of the paper do not necessarily reflect the position or the policy of the Defense Advanced Research Projects Agency. No official endorsement should be inferred.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 4

[2] S. Ahn, M. Gorlatova, P. Naghizadeh, M. Chiang, and P. Mittal. Adaptive fog-based output security for augmented reality. In *Proceedings of ACM SIGCOMM VR/AR Network Workshop*, 2018. 2

[3] F. Berton, L. Hoyet, A.-H. Olivier, J. Bruneau, O. Le Meur, and J. Pettré. Eye-gaze activity in crowds: Impact of virtual reality and density. In *Proceedings of IEEE VR*, 2020. 2, 6

[4] F. Biocca, C. Owen, A. Tang, and C. Bohil. Attention issues in spatial information systems: Directing mobile users' visual attention using augmented reality. *Journal of Management Information Systems*, 23(4):163–184, 2007. 2

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 3

[6] E. Brain. Sudoku rules - strategies, solving techniques and tricks. https://sudoku.com/sudoku-rules/, 2024. 3

[7] S. Brauns and J. Tümler. Simulation of the field of view in AR and VR headsets. In *Proceedings of Springer HCII*, 2021. 2

[8] A. P. Burgoyne, J. S. Tsukahara, C. A. Mashburn, R. Pak, and R. W. Engle. Nature and measurement of attention control. *Journal of Experimental Psychology: General*, 152(8):2369, 2023. 4, 7

[9] L. Chen, T. W. Day, W. Tang, and N. W. John. Recent developments and future challenges in medical mixed reality. In *Proceedings of IEEE ISMAR*, 2017. 1

[10] Y. Chen, P. S. Holzman, and K. Nakayama. Visual and cognitive control of attention in smooth pursuit. *Progress in Brain Research*, 140:255–265, 2002. 5

[11] F. Dehais, A. Lafont, R. Roy, and S. Fairclough. A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in Neuroscience*, 14:268, 2020. 1, 2

[12] B. J. Dixon, M. J. Daly, H. Chan, A. D. Vescan, I. J. Witterick, and J. C. Irish. Surgeons blinded by enhanced navigation: The effect of augmented reality on attention. *Surgical Endoscopy*, 27:454–461, 2013. 1, 2

[13] B. J. Dixon, M. J. Daly, H. H. Chan, A. Vescan, I. J. Witterick, and J. C. Irish. Inattentional blindness increased with augmented reality surgical navigation. *American Journal of Rhinology & Allergy*, 28(5):433–437, 2014. 1, 2

[14] J. Drewes, S. Feder, and W. Einhäuser. Gaze during locomotion in virtual reality and the real world. *Frontiers in Neuroscience*, 15:656913, 2021. 2

[15] S. Eom, D. Sykes, S. Rahimpour, and M. Gorlatova. Neurolens: Augmented reality-based contextual guidance through surgical tool tracking in neurosurgery. In *Proceedings of IEEE ISMAR*, 2022. 1

[16] R. Eyraud, E. Zibetti, and T. Baccino. Allocation of visual attention while driving with simulated augmented reality. *Transportation Research Part F: Traffic Psychology and Behaviour*, 32:46–55, 2015. 2

[17] Y. Gao, Y. Liu, J.-M. Normand, G. Moreau, X. Gao, and Y. Wang. A study on differences in human perception between a real and an AR scene viewed in an OST-HMD. *Journal of the Society for Information Display*, 27(3):155–171, 2019. 2

[18] Y. Ghasemi, H. Jeong, S. H. Choi, K.-B. Park, and J. Y. Lee. Deep learning-based object detection in augmented reality: A systematic review. *Computers in Industry*, 139:103661, 2022. 4

[19] N. Halin, J. E. Marsh, A. Hellman, I. Hellström, and P. Sörqvist. A shield against distraction. *Journal of Applied Research in Memory and Cognition*, 3(1):31–36, 2014. 7

[20] A. R. Harrivel, D. H. Weissman, D. C. Noll, and S. J. Peltier. Monitoring attentional state with fNIRS. *Frontiers in Human Neuroscience*, 7:861, 2013. 1

[21] K. Holmqvist and R. Andersson. Eye tracking: A comprehensive guide to methods. *Paradigms and Measures*, 2017. 1

[22] J. Hsu. *Multiple Comparisons: Theory and Methods*. CRC Press, 1996. 6

[23] A. R. Hunt, J. Reuther, M. D. Hilchey, and R. M. Klein. The relationship between spatial attention and eye movements. *Processes of Visuospatial Attention and Working Memory*, pages 255–278, 2019. 1

[24] L. A. Jäger, S. Makowski, P. Prasse, S. Liehr, M. Seidler, and T. Scheffer. Deep Eyedentification: Biometric identification using micro-movements of the eye. In *Proceedings of Springer ECML PKDD*, 2019. 2

[25] Y.-J. Jo, J.-S. Choi, J. Kim, H.-J. Kim, and S.-Y. Moon. Virtual reality (VR) simulation and augmented reality (AR) navigation in orthognathic surgery: A case report. *Applied Sciences*, 11(12):5673, 2021. 4

[26] H. Kim and J. L. Gabbard. Assessing distraction potential of augmented reality head-up displays for vehicle drivers. *Human Factors*, 64(5):852–865, 2022. 1, 2

[27] J. Kim, L. Hwang, S. Kwon, and S. Lee. Change in blink rate in the metaverse VR HMD and AR glasses environment. *International Journal of Environmental Research and Public Health*, 19(14):8551, 2022. 2

[28] J. Kim, Y. Sunil Kumar, J. Yoo, and S. Kwon. Change of blink rate in viewing virtual reality with HMD. *Symmetry*, 10(9):400, 2018. 2

[29] E. I. Knudsen. Neural circuits that mediate selective attention: A comparative perspective. *Trends in Neurosciences*, 41(11):789–805, 2018. 1

[30] Z. J. Kong, Q. Xu, J. Meng, and Y. C. Hu. Accumo: Accuracy-centric multitask offloading in edge-assisted mobile augmented reality. In *Proceedings of ACM MobiCom*, 2023. 2

[31] L. V. Kulke, J. Atkinson, and O. Braddick. Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking. *Frontiers in Human Neuroscience*, 10:592, 2016. 1

[32] J. Lacoche, E. Villain, and A. Foulonneau. Evaluating usability and user experience of AR applications in VR simulation. *Frontiers in Virtual Reality*, 3:881318, 2022. 1, 2

[33] G. Lan, B. Heit, T. Scargill, and M. Gorlatova. GazeGraph: Graph-based few-shot cognitive context sensing from human visual behavior. In *Proceedings of ACM SenSys*, 2020. 2

[34] K. Lebeck, K. Ruth, T. Kohno, and F. Roesner. Arya: Operating system support for securely augmenting reality. *IEEE Security & Privacy*, 16(1):44–53, 2018. 2

[35] C. Lee, G. A. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman. The effects of visual realism on search tasks in mixed reality simulation. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):547–556, 2013. 1, 2

[36] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[37] J.-C. Liu, K.-A. Li, S.-L. Yeh, and S.-Y. Chien. Assessing perceptual load and cognitive load by fixation-related information of eye movements. *Sensors*, 22(3):1187, 2022. 6

[38] L. Liu, H. Li, and M. Gruteser. Edge assisted real-time object detection for mobile augmented reality. In *Proceedings of ACM MobiCom*, 2019. 2

[39] D. Lohr, S. Aziz, L. Friedman, and O. V. Komogortsev. GazeBaseVR, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality. *Scientific Data*, 10(1):177, 2023. 7

[40] D. Lohr and O. V. Komogortsev. Eye Know You Too: Toward viable end-to-end eye movement biometrics for user authentication. *IEEE Transactions on Information Forensics and Security*, 17:3151–3164, 2022. 2

[41] S. Makowski, P. Prasse, D. R. Reich, D. Krakowczyk, L. A. Jäger, and T. Scheffer. DeepEyedentificationLive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):506–518, 2021. 2

[42] R. W. Marklin Jr, A. M. Toll, E. H. Bauman, and J. J. Simmins. Effect of two common head-mounted augmented reality systems on mus-

cle force and blink rate of electric utility power plant operators. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 1132–1136. SAGE Publishing, 2019. 2

[43] B. Massé, S. Ba, and R. Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2711–2724, 2017. 1

[44] V. Mazzia, F. Salvetti, and M. Chiaberge. Efficient-CapsNet: Capsule network with self-attention routing. *Scientific Reports*, 11(1):14634, 2021. 3

[45] A. M. Michalek, G. Jayawardena, and S. Jayarathna. Predicting ADHD using eye gaze metrics indexing working memory capacity. In *Computational Models for Biomedical Reasoning and Problem Solving*, pages 66–88. IGI Global, 2019. 7

[46] A. Olsen. The Tobii I-VT fixation filter. *Tobii Technology*, 21:4–19, 2012. 5

[47] A. Plopski, T. Hirzle, N. Norouzi, L. Qian, G. Bruder, and T. Langlotz. The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality. *ACM Computing Surveys*, 55(3):1–39, 2022. 1, 8

[48] K. Poels, de Yaw Yvonne Kort, and W. W. IJsselsteijn. D3.3 : Game experience questionnaire: Development of a self-report measure to assess the psychological impact of digital games. Technische Universiteit Eindhoven, 2007. 4

[49] Puzzles.ca. Free printable Sudoku puzzles. https://www.puzzles.ca/sudoku/, 2024. 4

[50] E. Ragan, C. Wilkes, D. A. Bowman, and T. Hollerer. Simulation of augmented reality systems in purely virtual environments. In *Proceedings of IEEE VR*, 2009. 1, 2

[51] D. Ren, T. Goldschwendt, Y. Chang, and T. Höllerer. Evaluating wide-field-of-view augmented reality with mixed reality simulation. In *Proceedings of IEEE VR*. IEEE, 2016. 2

[52] P. Renner and T. Pfeiffer. Attention guiding techniques using peripheral vision and eye tracking for feedback in augmented-reality-based assistance systems. In *Proceedings of IEEE 3DUI*, 2017. 2

[53] P. Renner and T. Pfeiffer. Attention guiding using augmented reality in complex environments. In *Proceedings of IEEE VR*, 2018. 2

[54] K. M. Rischer, A. M. González-Roldán, P. Montoya, S. Gigl, F. Anton, and M. van der Meulen. Distraction from pain: The role of selective attention and pain catastrophizing. *European Journal of Pain*, 24(10):1880–1891, 2020. 7

[55] A. Rosebrock. OpenCV Sudoku solver and OCR. https://pyimagesearch.com/2020/08/10/opencv-sudoku-solver-and-ocr/, 2024. 3

[56] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of ACM ETRA Adjunct*, 2000. 5

[57] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. de Jesus Filipe. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In *Proceedings of Springer TECH-EDU*, 2022. 2

[58] T. Shotton and J. H. Kim. Assessing differences on eye fixations by attention levels in an assembly environment. In *Proceedings of Springer AHFE*, 2020. 1

[59] V. Skaramagkas, E. Ktistakis, D. Manousos, E. Kazantzaki, N. S. Tachos, E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis. eSEE-d: Emotional state estimation based on eye-tracking dataset. *Brain Sciences*, 13(4):589, 2023. 2

[60] P. Smith, M. Shah, and N. da Vitoria Lobo. Determining driver visual attention with one camera. *IEEE Transactions on Intelligent Transportation Systems*, 4(4):205–218, 2003. 1

[61] B. V. Syiem, R. M. Kelly, J. Goncalves, E. Velloso, and T. Dingler. Impact of task on attentional tunneling in handheld augmented reality. In *Proceedings of ACM CHI*, 2021. 1, 2

[62] P. Tarnowski, M. Kołodziej, A. Majkowski, R. J. Rak, et al. Eye-tracking analysis for emotion recognition. *Computational Intelligence and Neuroscience*, 2020, 2020. 2

[63] N. Tobisková, L. Malmsköld, and T. Pederson. Multimodal augmented reality and subtle guidance for industrial assembly – a survey and ideation method. In *Proceedings of Springer HCII*, 2022. 1

[64] M. Tonnis, C. Sandor, G. Klinker, C. Lange, and H. Bubb. Experimental evaluation of an augmented reality visualization for directing a car driver's attention. In *Proceedings of IEEE ISMAR*, 2005. 2

[65] D. Van Der Linden, M. Tops, and A. B. Bakker. The neuroscience of the flow state: Involvement of the locus coeruleus norepinephrine system. *Frontiers in Psychology*, 12:645498, 2021. 5

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 7

[67] M. Wang, V. Callaghan, J. Bernhardt, K. White, and A. Peña-Rios. Augmented reality in education and training: pedagogical approaches and illustrative case studies. *Journal of Ambient Intelligence and Humanized Computing*, 9:1391–1402, 2018. 1

[68] T. Wang, X. Qian, F. He, X. Hu, K. Huo, Y. Cao, and K. Ramani. CAPturAR: An augmented reality tool for authoring human-involved context-aware applications. In *Proceedings of ACM UIST*, 2020. 2

[69] S. C.-C. Weng, T. Hopkins, R. Vanukuru, C. Tobin, A. Banic, D. Leithinger, and E. Y.-L. Do. How field of view affects awareness of an avatar during a musical task in augmented reality. In *Proceedings of IEEE VRW*, 2023. 2

[70] C. D. Wickens and A. L. Alexander. Attentional tunneling and task management in synthetic vision displays. *The International Journal of Aviation Psychology*, 19(2):182–199, 2009. 1

[71] C. D. Wickens, J. S. McCarley, and R. S. Gutzwiller. *Applied Attention Theory*. CRC press, 2022. 8

[72] S. Yang, J. R. Carlson, and S. Chen. How augmented reality affects advertising effectiveness: The mediating effects of curiosity and attention toward the ad. *Journal of Retailing and Consumer Services*, 54:102020, 2020. 2

[73] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of ACM SIGKDD*, 2021. 2, 7

[74] L. J. Zheng, J. Mountstephens, and J. Teo. Four-class emotion classification in virtual reality using pupillometry. *Journal of Big Data*, 7:1–9, 2020. 2