Neighbor-Environment Observer: An Intelligent Agent for Immersive Working Companionship

Zhe Sun, Qixuan Liang, Meng Wang, Zhenliang Zhang*

{sunzhe,liangqixuan,wangmeng,zlzhang}@bigai.ai

National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI)
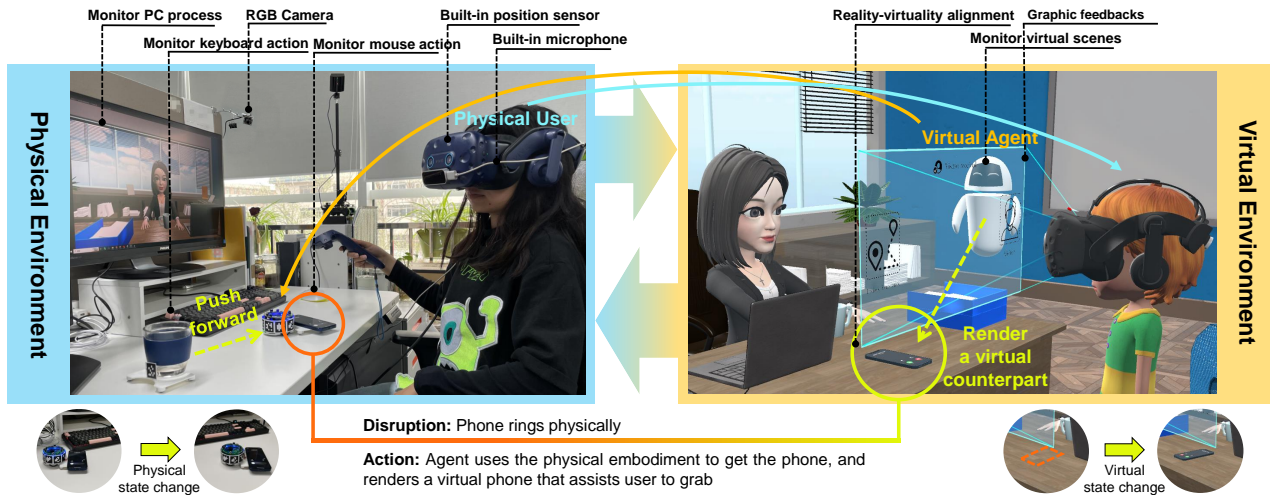
Beijing, China

Figure 1: The NEO system jointly observes the physical and virtual environments and takes action in the two environments simultaneously. It is designed to deal with disruptions for people when they are immersed in a virtual working environment.

ABSTRACT

Human-computer symbiosis is a crucial direction for the development of artificial intelligence. As intelligent systems become increasingly prevalent in our work and personal lives, it is important to develop strategies to support users across physical and virtual environments. While technological advances in personal digital devices, such as personal computers and virtual reality devices, can provide immersive experiences, they can also disrupt users' awareness of their surroundings and enhance the frustration caused by disturbances. In this paper, we propose a joint observation strategy for artificial agents to support users across virtual and physical environments. We introduce a prototype system, neighbor-environment observer (NEO), that utilizes non-invasive sensors to assist users in dealing with disruptions to their immersive experience. System experiments evaluate NEO from different perspectives and demonstrate the effectiveness of the joint observation strategy. A user study is conducted to evaluate its usability. The results show that

NEO could lessen users' workload with the learned user preference. We suggest that the proposed strategy can be applied to various smart home scenarios.

1 INTRODUCTION

Human-computer symbiosis is one of the potential outcomes of human-computer interaction (HCI). Licklider envisioned a partnership between humans and computers that could be viable, productive, and thriving [40]. Numerous efforts have been made from different perspectives to move closer to human-computer symbiosis, including computer vision [27, 33, 37, 60, 61], natural language

processing [2, 4, 11, 42], robotics [20, 34, 59], and interdisciplinary fields [13, 51, 52]. With the development of artificial intelligence (AI), remarkable achievements have emerged, such as deep learning (DL) [35] and reinforcement learning (RL) [55]. Intelligent agents now possess the capability to detect objects, recognize people's identities, comprehend sentences, and even offer human-like feedback through audio or visual avatars.

On the other hand, to approach an immersive working environment with digital content and agents, interface technologies such as virtual reality (VR), augmented reality (AR), and mixed reality (MR) have been developed. These technologies provide people with rich visual experiences and novel interactions [29]. When people are immersed in virtual environments, their virtual states may change vastly. So do their physical states. In such cases, cooperating with humans requires awareness not only of their states in the virtual environments but also in the physical environment. This can benefit from a combination of digital AI agents and smart-home agents.

There has been extensive research exploring digital AI agents for VR, AR, and MR environments. One notable approach is embodied artificial intelligence [19], which is designed to learn and explore in virtual simulated environments. With virtual embodiments, agents explore the virtual environment using not only visual information but also knowledge gained from interaction. This enables them to perform various tasks in VR, from point navigation [18] to simple daily interactions [54]. Consequently, agents based on embodied AI are more suitable for cooperating with humans by modeling human users' minds, which has further catalyzed research in the area of multi-agent systems [28, 53].

However, human users do not solely stay in virtual environments, even when using VR Head-Mounted Devices (HMDs). They often transition between the physical and virtual environments or receive information from both simultaneously. For example, one still hears sounds and smells scents from the physical room around them when using VR HMDs. This phenomenon of transition can be a source of disruption and can even lead to VR sickness [10]. External stimuli, such as sudden temperature changes or gusts of wind, can disrupt immersion in the virtual world [57], as can unexpected visitors at the door. These changes in physical states are often not considered in traditional immersive systems but can be observed and intervened by an intelligent agent equipped with sensors and actuators. In this case, here comes the research question:

RQ: How can an artificial agent observe, understand, and cooperate with humans when they are immersed in the virtual environment, particularly in the presence of disruptions from the physical world?

Researchers have been examining the question of how do virtual agents effectively operate in both physical and virtual environments, and one perspective is the concept of symmetrical reality (SR), as proposed by Zhang et al. [65–68]. The SR paradigm involves enabling virtual agents to perceive and act in both environments. Our work inherits the core concepts of SR paradigm and extends it in terms of technical implementation, prototype frameworks, and application cases. An example is shown in Figure 1. In particular, we define the physical and virtual environments as "neighbor environments", as any digital device can serve as a gateway between

them. We formulated the following research questions to further delineate the overarching research question *RQ*:

RQ-1: What methodology should be employed for an agent to understand users' states when they are in physical and virtual environments simultaneously?

RQ-2: How can an agent align and parse information from both users and the environments in a coherent manner, and generate decisions autonomously without explicit commands from the users?

RQ-3: What types of embodiments are necessary for the agent to execute decisions that involve both the physical and virtual environment?

RQ-4: What design strategy of the agent's interface could reduce the users' disruption and enhance immersion?

These research questions guide our investigation into the underlying factors that are critical to the development of an effective and efficient agent that can seamlessly operate in both physical and virtual environments. By addressing these questions, we hope to advance the state-of-the-art in human-agent interaction and facilitate the design of novel and innovative interfaces that can effectively serve users in complex and dynamic environments.

We proposed an intelligent agent, the neighbor-environment observer (NEO), as shown in Figure 2, which is designed to operate across the physical and virtual environments, accompanying human users in their daily activities. By parsing data from a non-invasive sensor system, NEO leverages a scalable decision module to generate real-time decisions that align with user needs without the need for explicit user commands. We also provide a series of physical and virtual embodiments to take action across environments. We then conducted a user study to assess the usability of NEO. The results demonstrate that NEO effectively reduces user workload resulting from disruptions during VR immersion, increasing user engagement. These findings highlight the importance of developing agents that can seamlessly operate across physical and virtual environments, providing personalized and responsive support to users in dynamic and complex settings.

In summary, this paper makes four key contributions.

- We design a framework for neighbor-environment observation that leverages non-invasive sensors to gather information from both physical and virtual environments.
- We propose a decision-making algorithm that can (1) generate decisions autonomously, (2) is compatible with different levels of sensor implementation, and (3) can learn user preferences interactively.
- We propose a series of physical and virtual embodiments and a matching method so that they can collaborate to execute the generated decisions across both environments.
- We develop a prototype system, NEO, as an artificial secretary in a smart home scenario, specifically a study room.

2 RELATED WORKS

2.1 Environment Observation

Multiple sensors have been implemented in smart spaces to gain a better perception of the physical environment. Cameras and microphones are widely used across research and industrial fields. RGB
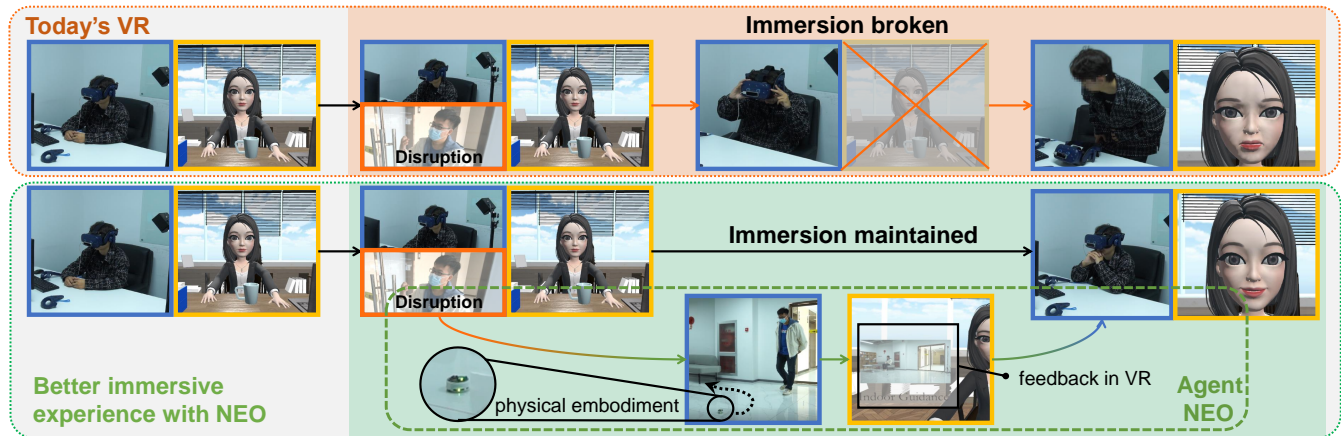
Figure 2: External stimuli and disruptions may break the immersion. We propose an intelligent agent NEO that could perceive user's demands, help them to handle the disruptions, and preserve their immersion in virtual environments.

cameras are used for 2D object detection, while depth sensors are usually used for 3D object detection and robotic navigation [21]. Thermal sensors are used to monitor temperature, which can also be used as safety alarms. Dowdall et al. [15] proposed a system to support the life of senior people. They used burglar alarm-style sensors to detect the states of doors and windows. A custom-made body count sensor was used to detect the number of people in the environment. Kidd's team [30] proposed the design of a system using small radio-frequency tags to track the movement of objects and show their locations. Borelli et al. [6] proposed an extensive digital platform for smart homes named HABITAT with the technology of the Internet of Things (IoT). They utilized radio frequency identification (RFID) readers and tags for localization.

2.2 User Understanding

Understanding human behavior and intention is a widely discussed topic. Thakur et al. [58] used skeleton data collected by wearable sensors to analyze interactions of the aged group in smart homes. They infer emergencies by a rule-based reasoning diagram.

Attention, sometimes also denoted as engagement [56], is another important criterion studied in smart-home-related research. Data is collected through different methods including questionnaires [43], RGB or depth cameras, temporal performances, interviews, observations, physiological sensors (e.g., electroencephalography), tracking sensors (e.g., motion, eye and laser tracking), speech and dialogue records, contextual and application record (e.g., game score), etc. Visual features such as head pose and eye gaze are widely used as a reference for attention understanding [12, 45, 47]. The algorithm may overlook evidence of distraction when only using one feature. For example, when the user stares at what he/she is supposed to focus on, it is difficult to detect the state of daze [45]. Thus, some researchers turn to fusing modalities for better solutions [5, 62]. Rudovic et al. [48], for example, fused data from an RGB camera, a microphone, and a smart watch [1] to evaluate children's attention during child-robot interactions.

There has been a significant amount of work focused on the detection of interruptibility [26, 32] using microphones [16], personal computer (PC) activity [25], and motion sensors [23].

2.3 Embodied Agent in Virtual Environments

It is a rising topic that gives AI agents virtual embodiments and enables them to explore the virtual world. There are virtual (or physical-virtual mixed) environments designed for agents to learn skills from human demonstration [9, 38, 50]. Researchers also developed virtual simulators where AI agents learn new skills by interacting with the environments with their embodiments, without human demonstration. For example, CARLA [14], AI2-THOR [31], Virtual Home [44], iGibson [39, 63] are popular simulators. Therefore, we believe that the embodiments are necessary to build an agent for immersive working companionship. We decided to extend the design of embodiments to the physical environment and propose a matching method enabling their collaboration.

2.4 Agent in Smart Spaces

Some systems provide feedback by tuning the environment [46, 69] while others utilize avatars with digital cartoon faces [5, 12], simplified robotic bodies [17, 56], or humanoid bodies [24, 48].

Danninger et al. [12] proposed a virtual secretary system in the office that mediated phone calls and visits by implementing an RGB camera and a microphone in the office. Zhao et al. [69] and Richer et al. [46] designed a system that changed its appearance according to the users' physical states to help them keep concentrated or recover from stress. Cameras and wearable sensors were used to understand users' states while projectors, lighting systems, and sound systems were used to control the room's appearance.

3 NEIGHBOR-ENVIRONMENT OBSERVATION

3.1 Concept

Symmetrical Reality. *Symmetrical Reality* (SR) is a framework that unifies concepts and interactions about the combination of the physical world and the virtual world [67]. The two worlds are topologically symmetrical. The virtual world is composed of

---

virtual agents and the virtual environment, similar to how the physical world is composed of humans and the environment. The virtual environment and virtual agents can evolve automatically, similar to how the physical environment and humans can evolve. SR defines a symmetrical topology between the physical and the virtual world, highlighting the importance of considering virtual agents and human users at an equal level.

Neighbor-Environment Observation. In this article, we define the physical and the virtual environments as neighbor environments. Following the SR paradigm, neighbor-environment observation is an approach to fuse information from these two types of environments. As human users could interact in both environments, we suggest that only by observing the two environments simultaneously could the agent gain an overall knowledge of the ongoing situation. Therefore, such methodology enables the agent to better parse the states of the users and the environments. Here we consider the mapping between the physical environment and the virtual environment as a bijection. Cases, where one physical environment is linked to multiple virtual environments, will be discussed in the discussion chapter.

3.2 Requirements

When users immerse themselves in a virtual environment through digital devices, their physical bodies continue to gather information from the physical environment around them. However, this information can often be detrimental to the immersive experience in virtual reality, as disruptions from the physical environment such as a ringing phone or unexpected guests can easily break the immersion. Users may be forced to leave the virtual environment to address them, which might interrupt an important task (e.g., a report at an international conference or a negotiation with the boss). We contend that intelligent agents in smart homes should be able to recognize such dilemmas and provide spontaneous assistance to alleviate these disruptions. To achieve this, the agent must possess advanced cognitive abilities. Here list the requirements:

Req-1: Perception. The agent should be able to accurately perceive and understand the immediate situation, including the user states and the disruptions. (Related to *RQ-1*)

Req-2: Reasoning. The agent should be able to reason about the user's availability to handle disruptions and determine whether it is appropriate to intervene. (Related to *RQ-2*)

Req-3: Action. The agent should be capable of executing actions in both the virtual and the physical environments to minimize disruptions for the user. (Related to *RQ-3*)

Req-4: Feedback. The agent should provide appropriate feedback about the occurrence of disruptions while minimizing negative impacts on their immersion. (Related to *RQ-4*)

The four requirements are related to the four RQs accordingly.

Req-1 demands the agent to take both the physical and the virtual environment into account. Thus, sensors should be implemented into the two environments. Collected data should be aligned with a minimal time delay. Moreover, it requires the sensors to be non-invasive so that they will not interfere with the user's primary tasks. Data transfer should be wireless so that the agent can perceive a larger physical region.

Req-2 requires the agent to parse both the physical states and the virtual states of the user, as well as the environmental states. The agent should be able to distinguish when is appropriate to help and when is unnecessary. It requires an understanding of the user's goals and priorities. As user preferences are various, the agent's mind should be compatible with different users and should be able to update itself.

Req-3 requires the agent to have the ability to execute actions in the two environments. Sometimes it needs to manipulate objects in the physical world to handle certain disruptions. For example, when the battery is low, one should take off the VR HMD, find the charger, and plug it into the device. Neither graphical notifications nor voice assistants could solve such problems. Thus, we present a series of embodiments including physical ones like desktop robots and virtual ones such as notification user interfaces (UIs).

Req-4 demands the agent to understand the user's preferences and to provide clear and informative feedback. The feedback should be straightforward to alert the immersed user that something has happened and what has the agent done for the user. It should also be lightweight to avoid affecting others who are sharing the physical or virtual environment with the user. In addition, the agent should have a port to collect users' feedback and update itself accordingly.

4 METHOD

4.1 Design Overview

In this section, we propose a framework for intelligent agents to observe, understand and cooperate with humans when they immerse in a virtual environment. The framework is designed for single-user use. It is constructed by three modules: the perception module, the decision module, and the action module. Each module is designed with a specific purpose, utilizing distinct methods. We will provide a detailed explanation of these module individually.

Figure 3 shows an overview of how the information transfers between the three modules. Our framework is developed with the setting of common immersive work scenarios, utilizing a personal computer (PC) and a VR HMD. This allows quick evaluation of effectiveness with minimal use of customized sensors and devices. To increase portability, we employ wireless methods for data transfer.

4.2 Joint Observation

This method is focused on enabling the agent to access both physical and virtual environments, expanding the current capabilities of smart home agents, which primarily perceive the physical world. Physical sensors act as tentacles to gather information from the physical environment, while virtual data ports serve as tentacles to the virtual environments.

The collected information is grouped into four categories: (1) physical environmental information (PE); (2) physical user information (PU); (3) virtual environmental information (VE); and (4) virtual user information (VU). The sources included in this work are summarized in Table 1.

Physical environmental information. We add two additional sensors (two RGB cameras) to capture visual information in the physical environment. From this data, the agent detects events and disruptions that occur in the physical environment. The cameras'
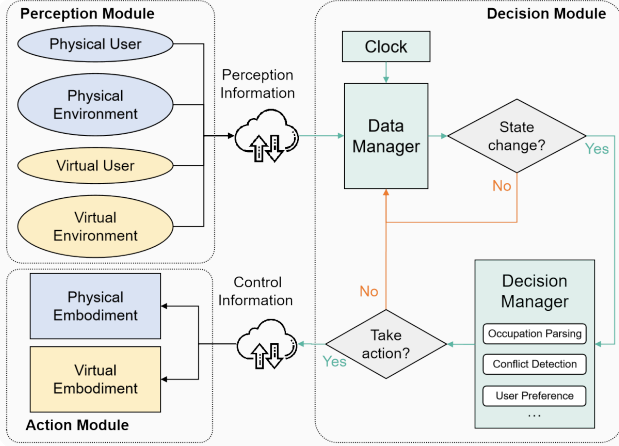
Figure 3: Information and decision transfer process in our framework. Joint observation information is transmitted to a data manager via WiFi. The data manager activates the decision manager, from which the decision is transmitted to the two types of embodiment.

positions define the agent's perception area. In the prototype system, as shown in Figure 1, one camera monitors the user's working area near the desk, while the other watches the door for visitors. The data from each camera is parsed independently. Our parsing algorithm does not require binocular-camera reconstruction, making it easy to change the number of cameras without calibration.

Physical user information. Visual and auditory information is utilized to interpret the user's states. An RGB camera is located beside the desk. The system classifies user poses into nine categories: using a device, using a keyboard, using a mouse, writing, reading, using a mobile device, resting, drinking, and eating. A microphone is used to detect whether the user is speaking. The system also infers three physiological needs by the time interval between the current time and the last detection of the corresponding action: thirst (activated after 30 minutes), hunger (activated after 180 minutes), and fatigue (activated after 2 hours of continuous work as prolonged periods may harm health [64], or after 20 minutes of VR usage considering visual fatigue [36]). Users can adjust these thresholds.

Virtual environmental information. The activated window on the PC desktop and the activated scene in VR are recorded. We define three kinds of mappings for PC windows: work (online meeting apps, browsers, Microsoft Office Word), entertainment (PC games, media players), and others. Similarly, we define three kinds of mappings for VR scenes: work (online meeting scenes), entertainment (VR games, painting Apps), and others. A thread is activated to monitor the battery levels and data from the controllers.

Virtual user information. We monitor four user activities: mouse usage, keyboard usage, VR HMD usage, and VR controller interactions. Mouse and keyboard activities pertain to interactions through the PC. The VR HMD position indicates the user's location in the virtual environment. With calibration, it also shows their physical location. The HMD's orientation shows the user's gaze direction. Additionally, input signals from the VR controllers are utilized to recognize the user's options, indicating whether the user is occupied or not.

Table 1: Information sources in joint observation

Information Category	Type of Sources[1]
Physical Environment (PE)	Visual information *(RGB camera)*.
Physical User (PU)	Visual information *(RGB camera)*; Auditory information *(Microphone)*.
Virtual Environment (VE)	Activated windows on PC; Activated VR scenes.
Virtual User (VU)	Activity of mouse and keyboard; Activity of VR HMD & controllers.

[1] Italics indicate that additional sensors are utilized to capture such information.

4.3 Scalable Decision Making

4.3.1 Decision Module.

The decision module is designed based on the information type from the perception module. It is initialized by an And-Or Graph (AOG) [70] with a probabilistic model. When receiving reports from the sensors, the decision module generates a parse tree by pruning the AOGs and categorizes possible user occupation states into two channels:

(1) **input channel:** hands occupation, speaking occupation.
(2) **output channel:** visual occupation, auditory occupation.

Joint parsing. To give a better vision of this model, we utilize the case where the user works from home using a VR HMD and PC. The AOG, shown in Figure 4, consists of two halves: the left half represents joint user observations ($G_u = (V_u, E_u)$), and the right half represents joint environment observations ($G_e = (V_e, E_e)$). The AOG contains And-nodes, Or-nodes, and Terminal-nodes [70]. Different layers represent different levels of observation. Terminal-nodes represent the occupation states to be observed.

Occupation inference. The decision module utilizes the parse tree generated from the AOG to infer potential occupations caused by disruptions and check for conflicts with existing occupations. The Terminal-nodes of the parse tree vote for the overall occupation. We introduce a hyperparameter to alter the weight between positive votes and negative votes. The agent infers the overall occupation state according to the percentage of positive votes among all votes.

Probabilistic model. The probability of taking actions at time t is denoted as Equation 1.

$$P(A_t, C_t, pt_t) = P(A_t|C_t, pt_t)P(C_t|pt_t)P(pt_t) \qquad (1)$$

where A_t denotes the decided action at time t, C_t denotes the conflict at time t, and pt_t denotes the parse tree at time t.

4.3.2 Personalization.

We propose a personalization module to adjust the probabilities of taking actions so that the agent learns the personal preferences of different users. We utilize a sigmoid function denoted as Equation 2 when tackling $P(A_t|C_t, pt_t)$.

$$sigmoid(x) = \frac{1}{1 + e^{-z}} \qquad (2)$$

For each $P(A|C, pt)$, we define an initial probability of taking action. After each decision the agent made, the user gives feedback to the agent for their action. According to the feedback, the conditional probability of taking action A_t oscillates between 0 and 1 along the sigmoid function. Figure 5 shows an example. Point M shows the initial conditional probability. If the agent receives negative
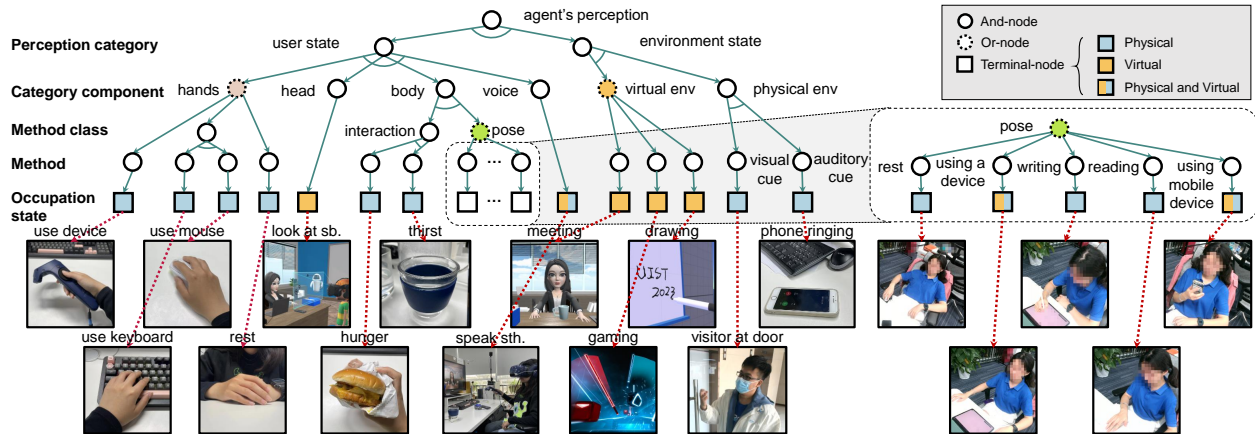
Figure 4: The And-Or Graph of the joint observation. The perception of the agent is categorized into two sets: user state and environment state. The user state node consists of four components, representing occupation methods, which are further divided based on method classes and specific methods. The environment state node consists of two components: the virtual part and the physical part. Colors of the Terminal-nodes denote whether the nodes belong to the physical environment or not.
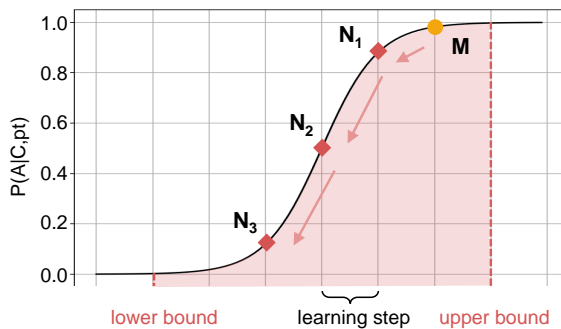

Figure 5: The agent adjusts the conditional probability along the sigmoid curve according to the user's feedback. It stops when reaching the upper or lower bound.

feedback, the probability takes one step left to reach point N_1 and takes $sigmoid(N_1)$ as the new conditional probability, and vice versa. If the user gives negative feedback again, the probability goes to point N_2 to update $P(A|C, pt)$.

4.4 Joint Action

To take action and offer assistance in the two environments, the agent needs different kinds of embodiment. We define the actions into four categories. **Graphical actions** in the virtual environment are shown as UI notifications, including (1) a reminder of taking a break; (2) a video stream of the visitor at the door; (3) a reminder of the devices' low battery state. **Virtual embodied actions** are designed to provide visual cues in the virtual environment. A robot character is shown if NEO starts monitoring. When bringing objects for the user, NEO renders their virtual counterparts for the user to grab the delivered gadgets without removing VR HMD. **Physical embodied actions** are designed for the agent to interact with the physical environment. The agent brings desktop-level objects and receives visitors. Details are introduced in the next chapter. **Auditory actions** play an auxiliary role in receiving visitors.
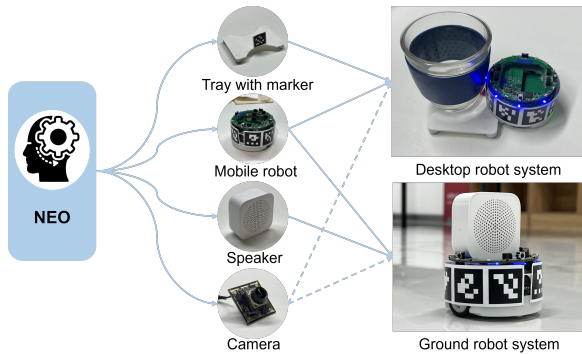
5 IMPLEMENTATION

Following these methods, we develop a prototype system, NEO.

5.1 Software Overview

Algorithm implementation. To simplify the system, similar algorithms are used to process data from similar sensor types (See Table 1 for sensors of NEO). For instance, images from RGB cameras are processed by similar algorithms developed based on CVZone [1], OpenCV [7] and MediaPipe [41]. Algorithms for cameras of the physical environment detect and crop region that includes faces in the image to detect visitors and inform the user. Algorithms for cameras of the physical user estimate the user's skeleton and detect objects near the user to categorize the pose. Virtual activities data of the user are smoothed by a filter before passing to the decision module. Other data is transferred to the decision module directly.

Interface implementation. In order to facilitate NEO's graphical actions, we design a series of interfaces. To show the design, we use the scenario of a one-on-one VR meeting as an example. As shown in Figure 1, NEO renders a robot character at startup to represent its running state to the user. NEO renders a virtual counterpart in the virtual environment when bringing an object to the user physically. By calibrating the position of NEO's physical embodiments, we align the position of physical objects and their virtual counterparts. For example, if NEO brings a cell phone to the user, a virtual phone is rendered in the VR scene, moving in sync with the physical phone. All virtual interface is only visible to users. Other individuals in the VR scene are unable to see the virtual objects and interfaces.

Data communication and management. A data manager module is developed to align the heterogeneous input data by a clock wave (T = 20 ms), as shown in Figure 3. The sensor states are checked at the clock-rising edge. If a sensor does not send data in the last clock period, its state will be set to *inactive*. To improve efficiency, the data manager module passes current states to the decision module only when a state change occurs.

Figure 6: Hardware implementation of NEO. The mobile robots serve as movable physical embodiments of NEO. Together with trays and the camera, they construct the desktop robot system that provides object transportation. The mobile robot connected with a speaker forms the ground robot system that could help the users to receive visitors.

5.2 Hardware Overview

Hardware implementation is shown in Figure 6.

Personal computer. The core computation of NEO is accomplished by the existing personal computer of the user. In this work, we implement NEO on a PC with an AMD Ryzen 9 5950X CPU and an NVIDIA RTX 3070 GPU. It is also used for rendering virtual environments for the experiment. Two 1080p RGB cameras are connected to the PC to capture visual information.

VR related devices. An HTC Vive HMD and two controllers are used. The HMD displays the virtual environment, while the controllers allow for interaction. They also provide VU sensors for NEO: the HMD provides the user's orientation and the built-in microphone collects auditory cues. A Leap Motion device is mounted in the front of the Vive HMD to reconstruct the skeleton of the user's hands for object grasping.

Movable embodiment. A desktop robot with a diameter of 7 cm is developed as the movable embodiment. The robot is equipped with two differential-driven wheels and two reduction stepper motors, by which it can move freely on the tabletop and locate itself by odometry. The linear speed of the robot can reach 10 cm/s with 3 cm diameter wheels, and the pull-out torque can reach 0.4 kgf/cm. With a wireless microcontroller unit (ESP32PICOV302) the robot is able to receive the message either from WiFi or from Bluetooth Low Energy (BLE) devices. Trays with markers and universal wheels (4 wheels for 1 tray) are 3D printed to support object transportation. The desktop robot could identify markers on the tray to locate the target objects on it. To bring objects to the user, the desktop robot moves to the marker, and pushes the tray toward the user.

5.3 VR-side Implementation

To help readers better understand our framework, we demonstrate it in a VR experience of a one-on-one meeting. Figure 7 shows the basic scene of the implementation, some examples of the disruptions, and the corresponding actions provided by NEO. During the meeting, the VR users discuss a research topic with the avatar of a professor. The professor introduces her point of view in the meeting and brings up several open questions. Different types of

disruptions, either from the physical environment or the virtual environment, are rendered during the meeting. The VR users are asked to concentrate on the meeting as much as possible while handling these disruptions.

6 EXPERIMENT

We launched two technical experiments to evaluate the ability of NEO from two perspectives: (1) evaluation #1: the influence of different activated sensor groups; (2) evaluation #2: NEO's ability to learn the user's preference.

6.1 System Evaluation #1

6.1.1 Experiment design.

Evaluation goal. Measure the extent of NEO's capability when activating different sensor groups, related to the *Req-1* and *Req-2*.

Apparatus. We prepared a toy test set comprising different online working or entertaining scenarios with two dimensions: occupation, and disruption. There are 13 cases in the occupation set (O01-O13) and 6 cases in the disruption set (D1-D6). O13 was included as it is common in daily life. Testing this case will make the system evaluation more comprehensive. Sensors were categorized into four classes according to their information source type listed in Table 1: physical environment (PE), physical user (PU), virtual environment (VE), and virtual user (VU). To make the result clearer, we chose 10 typical sensor combinations (S1-S10). Table 2 shows the details.

Task. We explored the system's ability by activating different sensor groups and testing the decision accuracy of NEO. The accuracy here is defined as Equation 3:

$$Acc = \frac{\sum_{d \in D} correct(d)}{N_{decision}} \quad (3)$$

where $N_{decision}$ is the total number of decisions in a certain case, D is the set of decisions from NEO in this case, and $correct(d)$ is defined as:

$$correct(d) = \begin{cases} 1 & , \quad d = d_{gt}, \\ 0 & , \quad else. \end{cases} \quad (4)$$

6.1.2 Result.

The result is shown in Figure 8. We visualized a total of 78 different cases with 10 different combinations of sensors. Overall, the results show an increasing trend in accuracy when activating more sensors. The performance on O13 reaches 100%, showing that NEO knows when not to help. The level S10 that has access to all four observation classes reaches the highest performance. S10 is able to handle most of the cases except for D6. All sensor level shows zero accuracy in this case (except for O13). It is because the current version of NEO fails to parse the disruption of "knock over a glass of water" so its decision remains "no action". Disruptions like D6 are denoted as hard problems. However, such shortcomings of NEO could be reduced in future work by extending its parsing algorithm.

For disruption case 1 (D1), we noticed that S7 failed to react to all the disruptions. This was caused by the missing perception of the physical environment. Information of D1, a visitor at the door, can only be observed in the physical environment. Losing connection to PE sensors caused this fatal problem. Similar phenomena were
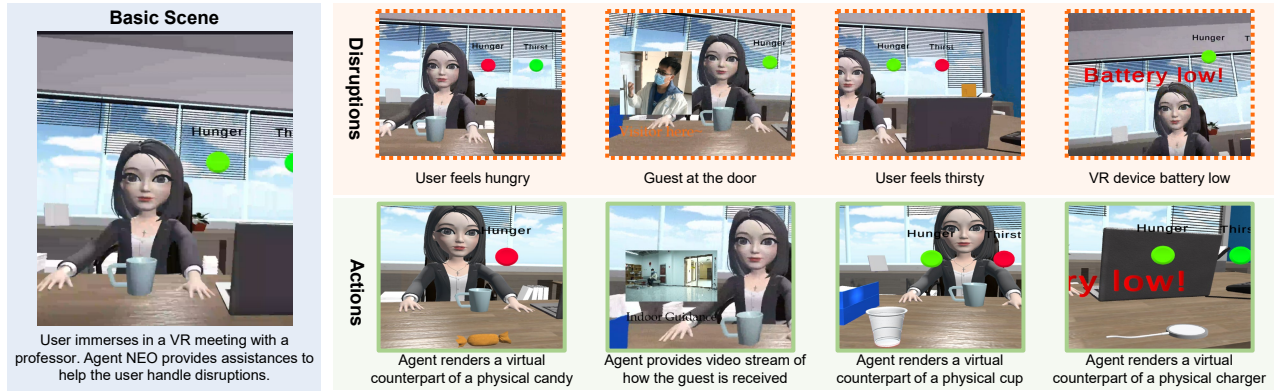
Figure 7: VR implementation of NEO. It visualizes how the disruptions and actions are rendered in VR.

Table 2: Dimensions and cases in the test set.

Activated Sensor Group		Occupation		Disruption	
ID	**Case**	**ID**	**Case**	**ID**	**Case**
S1	PE only	O1	Have a one-on-one discussion in VR	D1	Visitor knocks at the door
S2	VU only	O2	Play a game in VR	D2	User feels thirsty
S3	PU only	O3	Paint in VR with controllers	D3	User keeps using digital device over 20 minutes
S4	VU-VE	O4	Have a multi-person meeting in VR	D4	Phone rings
S5	PE-VU	O5	Watch a movie in VR	D5	VR device notifies battery low
S6	PU-PE	O6	Have an online interview using PC	D6	User knocks over a glass of water
S7	PU-VU-VE	O7	Play game on PC		
S8	PU-PE-VU	O8	Write an article on PC		
S9	PU-PE-VE	O9	Have an online lesson on PC		
S10	PU-PE-VU-VE	O10	Watch a movie on PC		
		O11	Use a mobile device (smartphone or tablet)		
		O12	Read a book		
		O13	Have a rest		

observed in D4 (S7), D5 (S8), etc. These phenomena support the theory of joint observation: losing connection to either the physical environment or the virtual environment will lead to a drop in performance.

O10 is a relatively demanding task where NEO's performance is sensitive to the information of VE. Clues from PU are similar to O13. Only by observing the virtual environment, namely the active window on the PC, can NEO understand the true state. This also supports the joint observation theory.

6.2 System Evaluation #2

6.2.1 Experiment design.

Evaluation goal. Measure the capability of NEO to learn user preferences while testing whether the system fulfills *Req-2*.

Apparatus. We chose a typical case (O1+D1) from the test set to evaluate the ability of personalization. The occupation case was set to always be "have a one-on-one discussion in VR" while the disruption was always "a visitor knocks at the door". We put three simulated users in the simulated environment with different preferences about NEO's actions, which is shown in Table 3. User A always gets annoyed if NEO goes and receives the visitor. User B gets annoyed in the first 3 rounds, shows neutral attitudes in the middle 4 rounds, and reacts satisfied in the last 3 rounds. User C is always satisfied with NEO's help.

Task. NEO was allowed to interact with the three users 10 times to learn their preferences.

6.2.2 Result.

The change in the action probability during the 10-round interactions is visualized in Figure 9. In the case of User A, the probability of receiving the visitor drops fast after the first three interactions. In the case of User B, a similar drop is seen in the first two rounds. The probability raises back after User B changes its mind at the last three interactions. In the case of User C, the action probability remains high as User C likes NEO to receive visitors. The result visualizes the learning procedure of NEO and demonstrates the ability of NEO to learn different preferences.

6.3 User Study

6.3.1 Goal and Hypothesis.

Evaluation goal. The goal of this user study is to validate whether our method can autonomously generate decisions when disruptions occur and reduce users' distraction through the embodiments. Moreover, we want to evaluate the usability of NEO. We invited the participants to come to our lab and experience a VR meeting. During the meeting, we rendered multiple disruptions around the participants to record their reactions and feedback.

Table 3: Personality of the simulated users.

User	Personality
User A	Always annoyed if NEO goes and receives the visitor.
User B	Annoyed at first, neutral in the middle, satisfied at last.
User C	Always satisfied if NEO goes and receives the visitor.
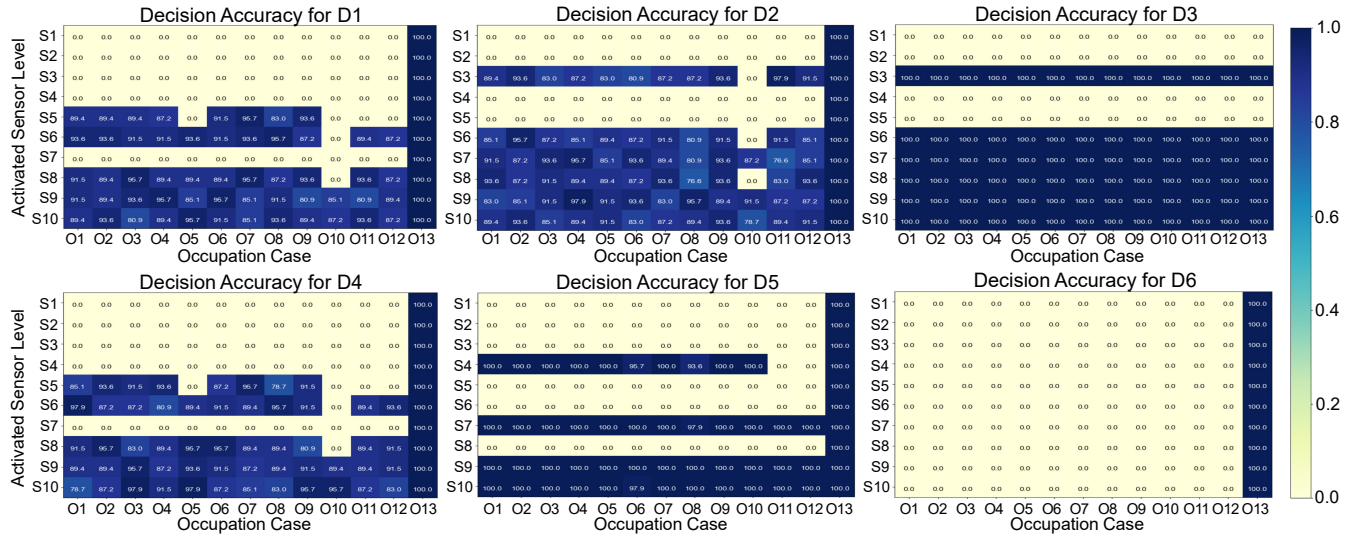
Figure 8: Result of system experiment. The six sub-graphs visualize the decision accuracy of NEO in the occurrence of the six disruption cases (D1-D6) respectively. The x-axis of each graph denotes the ID of the tested occupation case (O1-O13). The y-axes denote the ID of the activated sensor group (S1-S10). All details of the three groups of variables are shown in Table 2. Numbers in the matrix refer to accuracy in terms of percentage. The percent symbol is hidden for a clearer view.
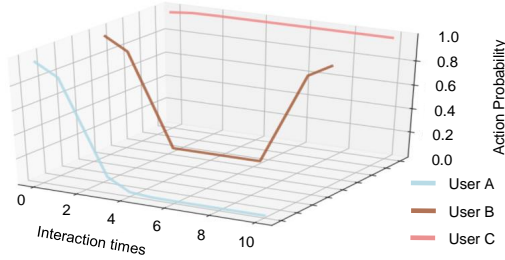
Figure 9: Result of learning experiment with three users.

Hypothesis. While the two system evaluations address *Req-1* and *Req-2*, the following hypotheses pertain to *Req-3* and *Req-4*.

(H1) NEO could enhance users' sense of engagement during a VR meeting in the presence of external disruptions. We hypothesize that NEO's assistance in neighboring environments thereby increases user engagement when disruptions occur. **(H2) NEO could alleviate users' workload in managing distracting events.** We hypothesize that when NEO shares a portion of the workload to disruptions, that of users is reduced. **(H3) The personalization module would improve the overall user experience of the system.** Although the action rate is designed based on occupation conflict, we acknowledge that people have varying preferences regarding NEO's assistance. Therefore, we postulate that user experience will be further enhanced after personalization.

6.3.2 Participants.

Twelve participants (P01-P12, 5 Females and 7 Males) aged between 21 to 31 were recruited (M=24.92, SD=3.20). About 83.3% of the participants have used VR devices and 66.7% work online over eight times per month, which is considered relatively frequent. As compensation for their effort, participants received a small gift. All

participants attended in person and reached the end of the user study.

6.3.3 Procedure.

The whole procedure lasted for 30 minutes. An investigator helped the participants to wear the VR HMD and took notes during the study. After reporting demographic information, participants were required to go through three phases of tasks: (1) the Control phase, (2) the Original NEO (O-NEO) phase, (3) the Personalized NEO (P-NEO) phase. If participants get familiar with the procedure, the ratings of engagement might naturally increase and the workload might drop. To exclude such influence, we disrupted the order of the phases (e.g., P01 went through Control phase → O-NEO → P-NEO while P02 went through P-NEO → O-NEO → Control phase).

The procedure is visualized in Figure 10. Participants were required to focus on a discussion with a professor through a VR meeting. Six disruptions occurred in turn during the meeting. The professor discussed different topics in different phases to exclude the influence of familiarity. The three phases shared this same basic track:

[00:00] Beginning of the experiment;
[00:30] Courier visiting;
[00:40] Professor starting a certain topic;
[01:10] User feeling hungry;
[01:40] Guest arriving;
[02:10] Cell phone ringing;
[02:40] Professor raising a question about the topic;
[03:00] User feeling thirsty;
[03:30] VR device battery low.

We rendered an alternative notice in the VR scene for physiological need, as shown in Figure 7. The participants were asked to pretend that they were hungry or thirsty when the corresponding buttons turned from green to red. We rendered the disruption of
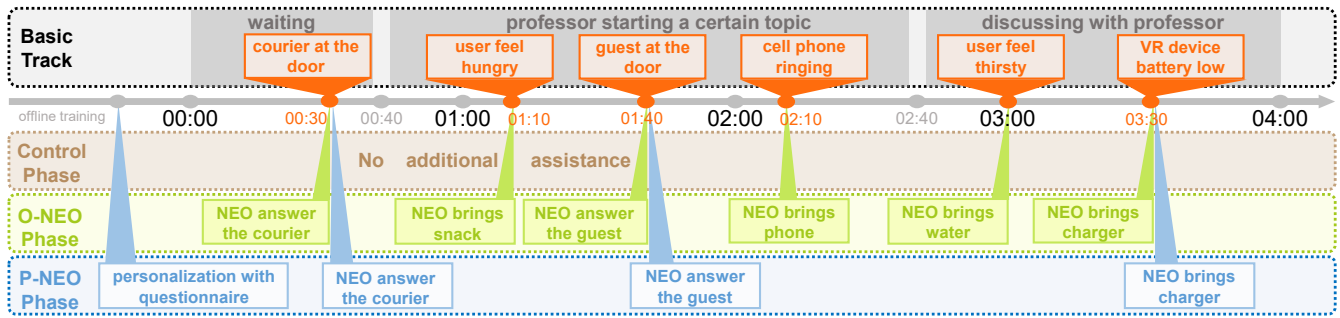
Figure 10: Procedure of the user study. The three phases shared the same basic track: the participant was asked to have a discussion with a professor. Several disruptions were rendered and the participant was required to handle them. Assistance in the P-NEO phase was different according to their preference. What is shown in this figure is an example of a possible case.
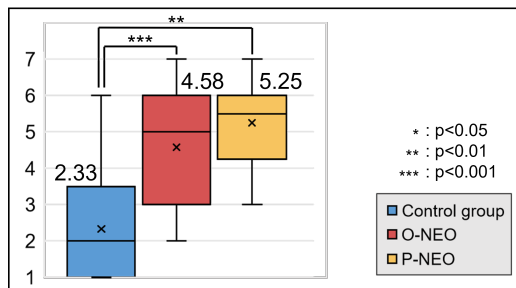
Figure 11: Ratings for engagement of VR meeting.

"battery low" similarly. Note that both the alternative notice and the virtual interfaces were exclusively visible to the participants.

Each phase had its unique parts as listed below:

Control phase: No additional assistance was provided.

O-NEO phase: In this phase, O-NEO helped the participants to deal with various disturbances regardless of their preferences.

P-NEO phase: Before the start of this phase, all participants were asked to complete a personalized questionnaire for NEO to learn their preferences. During this phase, P-NEO took action selectively.

At the end of each phase, participants filled in a post-study questionnaire. The post-study questionnaire consists of four parts: (1) a questionnaire asking participants to assess their engagement during the meeting [3, 8] (related to **H1**); (2) two NASA Task Load Index (NASA-TLX) questionnaires [22] to evaluate the overall workload of the whole phase and that of handling the disruptions (related to **H2**); (3) a short version of the user experience questionnaire (UEQ-S) [49] to quantify the user experience of O-NEO and P-NEO (related to **H3**); (4) a questionnaire to assess participants' satisfaction about O-NEO and P-NEO (related to **H3**). For consistency, all measures are on a 7-point Likert (1=very low, 7=very high).

A semi-structured interview was launched at last.

6.3.4 Result.

We utilized paired t-tests for significance analysis between every two phases if the data fit a normal distribution. We utilized the Wilcoxon signed rank tests when the normality assumption was violated.

Engagement of the meeting.

Figure 11 shows the reported engagement points. The average degree of engagement increases from 2.33 to 4.58 after the

O-NEO is involved. The average degree of engagement of the P-NEO phase is 0.67 higher than that of the O-NEO phase. A paired t-test shows a significant difference between the control phase and the O-NEO phase ($t_{11} = -4.700, p < 0.01$). Significance is also found in the difference between the control phase and the P-NEO phase ($t_{11} = -6.027, p < 0.001$). The result infers that NEO increases user engagement and helps the user to keep concentrating.

Work load.

The workload over the phase is shown in Figure 12(a). As most pairs of data in NASA-TLX violated the normality assumption, we utilized the Wilcoxon signed rank tests for significance analysis. The mental demand decreases in the O-NEO phase ($Z = -2.388, p = 0.017$) and the P-NEO phase ($Z = -2.701, p = 0.007$), compared to the control phase. The physical demand (PD) and the time demand (TD) decrease in the O-NEO phase (PD: $Z = -2.921, p = 0.003$, TD: $Z = -2.462, p = 0.014$) and the P-NEO phase (PD: $Z = -2.897, p = 0.004$, TD: $Z = -2.825, p = 0.005$) as well. The performance rating of the P-NEO phase is significantly higher than that of the control phase ($Z = -2.366, p = 0.018$). These two phases also show lower requirements of effort (O-NEO: $Z = -2.101, p = 0.036$, P-NEP: $Z = -2.589, p = 0.010$) and lower frustration levels (O-NEO: $Z = -2.573, p = 0.010$, P-NEO: $Z = -2.777, P = 0.005$) compared with the control phase. The difference between the O-NEO phase and the P-NEO phase fails to pass the significance analysis.

As for the workload of handling the disruptions (shown in Figure 12(b)), similar differences are observed by the Wilcoxon signed rank tests between the two pairs: (1) the control phase vs. the O-NEO phase, (2) the control phase vs. the P-NEO phase. Additional differences occur when comparing the effort and frustration levels between the O-NEO phase and the P-NEO phase. The effort required in the P-NEO phase is lower than that of the O-NEO phase ($Z = -2.271, p = 0.023$). The frustration level is also lower in the P-NEO phase ($Z = -2.333, p = 0.020$). The mean NASA-TLX score of the O-NEO phase and the P-NEO phase is reported to be 2.90 and 2.56 on an equal-weighted 7-point scale, indicating that the general workload of handling disruptions with the help of NEO is relatively light. These results imply that NEO has successfully cooperated with participants while they are occupied by something. The P-NEO with personalized knowledge can better release participants from handling the disruptions than the O-NEO, which shows the value of personalization.
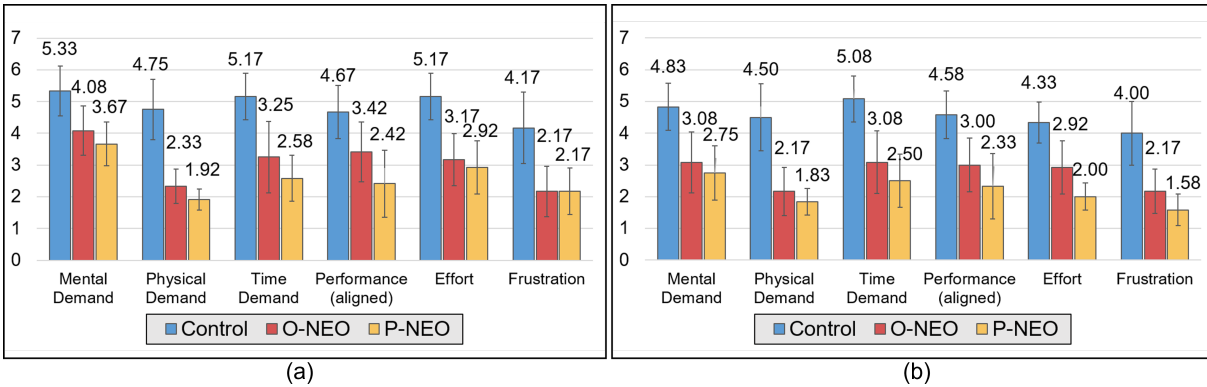
(a)   (b)

Figure 12: NASA-TLX results. (a) The overall workload of the whole task (including both discussing with the professor and handling disruptions). (b) The workload of handling the disruptions. Note that the points of performance in this figure are aligned so that all questions share the same trend: "the lower the point is, the better". (*shownpoint = 7 − originalpoint*)
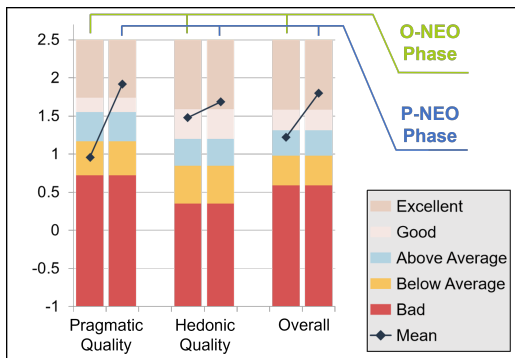

Figure 13: Results of the UEQ-S. In each group of bars, the left one indicates the result of O-NEO while the right one indicates the result of P-NEO.

Figure 14: Satisfaction ratings of NEO.

User experience.

The results of UEQ-S are shown in Figure 13. The overall ratings are both above average, indicating that NEO has provided participants with a good experience of cooperation. A significant improvement is captured in the channel of pragmatic quality (paired t-test, $t_{11} = -2.615, p = 0.024$) after personalization. We noticed that the average hedonic quality went up a level after personalization: from good to excellent. The average of P-NEO's overall ratings (excellent) is two levels higher than that of O-NEO (good).

The average ratings of satisfaction that participants have reported for O-NEO is 5.25 (SD=1.06), while that for P-NEO reaches 6.00 (SD=0.74). Figure 14 visualizes the results. The ratings for NEO significantly increase after personalization (Wilcoxon signed rank test, $Z = -2.041, p = 0.041$).

Semi-structured interview.

Participants were asked to rate their will of using NEO in their daily lives on a 7-point scale (1=not at all, 7=very much so). The average rating is 6.25 (SD=0.75), showing that participants hold a positive attitude towards NEO. Participants' opinions about the design of NEO are summarized below.

(1) *The way NEO observes and acts.* The majority of the participants were satisfied with NEO. "It can really reduce the disturbance," said P07. "The design is simple but clever," agreed by P11, "and it
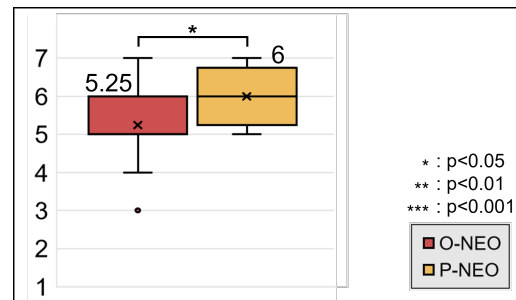
is pretty enough for me." Four expressed particular interest in the personalization function (P01, P06, P08-09). P09 and P10 mentioned that answering the door was an important requirement and had been well handled by NEO. P08 expressed the expectation of NEO to evaluate his attention and to offer water or food based on this evaluation: "I don't want water if I'm in a debate, but I might want some when chatting." P09 emphasized the importance of personalization. "I might not use NEO without personalization." said P09, "I don't want a model designed by strangers."

(2) *Other embodiment for NEO.* Four participants reported that the proposed embodiments were enough. P01, P09, and P12 would like to have additional robotic arms to carry large objects or pick up books from a book shell. Humanoid robots were nominated by two participants, while the idea of using drones was proposed by P06. Different from other participants, P11, a student who majored in robotics, expressed his concerns: "I think current embodiments are enough. Using other embodiment might cause additional danger."

(3) *More kinds of behavior.* P03 and P06 both suggested extending the audio module of NEO so that NEO could not only receive the visitors but also chat with them. The requirement of displaying caller information was addressed by P07, P09, and P12. P02, P03, P06, and P07 all pointed out the need for assistance in taking notes or searching for information during the meeting.

(4) *Ethical concerns.* P05 expressed worries about the delay in personalization. "I am afraid that NEO might open the door for my ex after we break up." said P05, "When I interact with NEO, it seems

like I'm looking at my past self." P06 reminded us that there might be issues if NEO observes confidential conversations.

7 DISCUSSION AND FUTURE WORK

7.1 System development

The meanings of joint observation. It is seen from Figure 8 that for most cases, the accuracies increase when more classes of sensors are activated. We thus stress the necessity of joint observation for an agent to cooperate with people in both physical and virtual environments. However, we also observe the difference in accuracy across cases. There are cases that reported good accuracy results with only PU sensors, but poor accuracy with only PE or VE sensors. This is caused by the difference in the number of cases that different sensor classes are related to. For example, the PU class contains a rich amount of information that related to the user's occupation state and disruption states. The VE class's information, on the other side, is related to fewer states than that of PU. Although involving more cases in the test set could reduce such differences, we argue that the listed cases already reflect the advantage of joint observation. There are also some cases that are not parsed by NEO, such as "knocking over a glass of water". This is caused by the limitation of NEO's current sensor set and could be further explored by extending sensor types for NEO.

Compatibility The results of Stage-2 demonstrate NEO's compatibility of cooperating with different kinds of people. Moreover, the results of Stage-1 show the compatibility of NEO to work with different sensor settings. This means that users can activate a subset of NEO's sensor set based on their needs and have NEO working without additional development. It shows NEO's potential of being implemented in conventional rooms. The users do not have to implement the whole system at the beginning.

7.2 User experience

Human-agent symbiosis. In the user study, we put our participants in a hectic setting where they must handle several things at once. Then, we contrasted their workload and level of involvement with and without NEO's assistance. It is concluded from the results that NEO did notice the disruption, made the right decision, and reduced the distraction through bi-environmental action. The feedback from participants through the UEQ-S and the semi-structured interview also suggests that NEO is helpful during the VR meeting, especially after personalization. Regarding the two physiological disruptions, although NEO has brought water and food to the user upon detection of thirst and hunger, saving their effort to take off the HMD, drinking and eating are still distracting to the immersive experience. We conclude from the above evidence that NEO takes one more step towards the two expectations of a man-computer symbiosis [40] to provide users with a not only viable but also productive partnership. The interview result also implies that NEO is well accepted among the participants. These conclusions picture a bright future of implementing NEO into homes as an AI partner.

Personalization. From another perspective, the results of UEQ-S show that the personalized NEO is preferable to the original one. The difference between the UEQ-S result of O-NEO and P-NEO supports the necessity of personalization. P09 also stressed its importance in the interview. Combining these with reports from the

interview that the P-NEO is more popular among the participants, we conclude that NEO did learn something from the participants during the simple feedback interaction session. The overall user experience of P-NEO is excellent, demonstrating the effect of personalization. However, as mentioned by P05, there is a delay between the feedback session and the actual change of behaviour. This delay may cause issues when sudden changes in user preferences happen.

7.3 Application

Systems like NEO could be implemented in more smart home scenarios. For example, it could be used in kitchens to give warnings when the food is about to burn; it could also detect pets if they enter the working area and warn the user, with extensions in the observation module. NEO could also be used in other scenarios such as health care. Disabled patients may immerse in virtual worlds for entertainment in the future. In this case, NEO could help maintain their physiological needs (bring food or medicines) or ensure their medical devices are running properly (notify medical staff when an infusion completes). Additionally, in manufacturing scenarios, if an accident occurs in a factory but unnoticed by humans, NEO could detect it through joint observation and handle the fault point.

7.4 Future work

With the capability and usability of NEO proved by these experiments, we are confident to integrate more sensors and algorithms into the system. Following the suggestions given by the participants, we would like to update NEO's set of embodiments with robotic arms and grippers in the future. The method of providing visitors information could be further improved to better align with the ongoing virtual scene.

We address the case of one physical environment with one virtual environment in this work. One physical environment corresponding to multiple virtual environments is a special case in SR. If there is only one user, NEO could handle these cases when the user switches from one virtual environment to another. If there are multiple users immersed in multiple virtual environments, NEO should be linked to each environment, and the collected data will be processed in the backend and integrated by the user. Such questions should be addressed in future work.

8 CONCLUSION

In this paper, we discuss methods for designing an agent under the topic of human-computer symbiosis. We propose a method that consists of three points: joint observation, scalable decision, and joint action. To verify the feasibility of this method, we develop a prototype system named NEO, which cooperates with people while they are immersed in a virtual environment. Two system experiments and a user study are conducted to evaluate this system. The results support the idea of joint observation and joint action. They also demonstrate the usability and effectiveness of NEO. Overall, NEO represents a promising step forward in the development of agents that can effectively operate across physical and virtual environments, facilitating human-agent interaction and providing users with personalized and responsive support, with important implications for future research in this field.

REFERENCES

[1] 2022. cvzone, cvzone: This is a Computer vision package that makes it easy to run Image processing and AI functions. At the core it uses OpenCV and Mediapipe libraries. https://github.com/cvzone/cvzone.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics* 11 (2019), 815–828.

[4] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems* 13 (2000).

[5] Jonnathan Berrezueta-Guzman, Ivan Pau, María-Luisa Martín-Ruiz, and Nuria Máximo-Bocanegra. 2020. Smart-home environment to support homework activities for children. *IEEE Access* 8 (2020), 160251–160267.

[6] Elena Borelli, Giacomo Paolini, Francesco Antoniazzi, Marina Barbiroli, Francesca Benassi, Federico Chesani, Lorenzo Chiari, Massimiliano Fantini, Franco Fuschini, Andrea Galassi, et al. 2019. HABITAT: An IoT solution for independent elderly. *Sensors* 19, 5 (2019), 1258.

[7] Gary Bradski. 2000. The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25, 11 (2000), 120–123.

[8] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing* 10, 4 (2017), 484–497.

[9] Moohyun Cha, Soonhung Han, Jaikyung Lee, and Byungil Choi. 2012. A virtual reality based fire training simulator integrated with fire dynamics data. *Fire Safety Journal* 50 (2012), 12–24.

[10] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. 2020. Virtual reality sickness: a review of causes and measurements. *International Journal of Human–Computer Interaction* 36, 17 (2020), 1658–1682.

[11] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, ARTICLE (2011), 2493–2537.

[12] Maria Danninger and Rainer Stiefelhagen. 2008. A context-aware virtual secretary in a smart office environment. In *Proceedings of the 16th ACM international conference on Multimedia*. 529–538.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. *arXiv preprint arXiv:1711.03938* (2017).

[15] Alan Dowdall and Mark Perry. 2001. The millennium home: Domestic technology to support independent-living older people. In *Proceedings of the 1st Equator IRC Workshop*. Citeseer, 1–15.

[16] James Fogarty, Scott E Hudson, and Jennifer Lai. 2004. Examining the robustness of sensor-based statistical models of human interruptibility. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 207–214.

[17] Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. 2017. Automatically classifying user engagement for dynamic multi-party human–robot interaction. *International Journal of Social Robotics* 9, 5 (2017), 659–674.

[18] Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. 2019. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1022–1031.

[19] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications* 12, 1 (2021), 5721.

[20] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2555–2565.

[21] Albert Haque, Arnold Milstein, and Li Fei-Fei. 2020. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature* 585, 7824 (2020), 193–202.

[22] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[23] Joyce Ho and Stephen S Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 909–918.

[24] Matthias Hoppe, Beat Rossmy, Daniel Peter Neumann, Stephan Streuber, Albrecht Schmidt, and Tonja-Katrin Machulla. 2020. A human touch: Social touch increases the perceived human-likeness of agents in virtual reality. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. ACM, 1–11.

[25] Eric Horvitz and Johnson Apacible. 2003. Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces*. 20–27.

[26] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 257–264.

[27] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).

[28] Unnat Jain, Luca Weihs, Eric Kolve, Ali Farhadi, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. 2020. A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 471–490.

[29] Brennan Jones, Yaying Zhang, Priscilla NY Wong, and Sean Rintel. 2020. Vroom: virtual robot overlay for online meetings. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.

[30] Cory D Kidd, Robert Orr, Gregory D Abowd, Christopher G Atkeson, Irfan A Essa, Blair MacIntyre, Elizabeth Mynatt, Thad E Starner, and Wendy Newstetter. 1999. The aware home: A living laboratory for ubiquitous computing research. In *Cooperative Buildings: Integrating Information, Organizations, and Architecture*. Springer, 191–198.

[31] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474* (2017).

[32] Panu Korpipää, Jonna Häkkilä, Juha Kela, Sami Ronkainen, and Ilkka Känsälä. 2004. Utilising context ontology in mobile device application personalisation. In *Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia*. 133–140.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[34] Steven M LaValle. 2006. *Planning algorithms*. Cambridge university press.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[36] Sang Hyeok Lee, Martha Kim, Hyosun Kim, and Choul Yong Park. 2021. Visual fatigue induced by watching virtual reality device and the effect of anisometropia. *Ergonomics* 64, 12 (2021), 1522–1531.

[37] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. 2018. High Performance Visual Tracking With Siamese Region Proposal Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Changyang Li, Wei Liang, Chris Quigley, Yibiao Zhao, and Lap-Fai Yu. 2017. Earthquake safety training through virtual drills. *Proceedings of IEEE Transactions on Visualization & Computer Graph (TVCG)* 23, 4 (2017), 1275–1284.

[39] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. 2021. IGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks. *arXiv preprint arXiv:2108.03272* (2021).

[40] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[41] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[43] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7 (2020), 92.

[44] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8494–8502.

[45] Libo Qiao, Zongyi Han, Wei Wang, Linlin Li, and Ying Tong. 2021. A Review of Attention Detection in Online Learning. *Artificial Intelligence in Education and Teaching Assessment* (2021), 87–100.

[46] Robert Richer, Nan Zhao, Bjoern M Eskofier, and Joseph A Paradiso. 2020. Exploring smart agents for the interaction with multimodal mediated environments. *Multimodal Technologies and Interaction* 4, 2 (2020), 27.

[47] Ognjen Rudovic, Hae Won Park, John Busche, Björn Schuller, Cynthia Breazeal, and Rosalind W Picard. 2019. Personalized estimation of engagement from videos using active learning with deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 217–226.

[48] Ognjen Rudovic, Meiru Zhang, Bjorn Schuller, and Rosalind Picard. 2019. Multimodal active learning from human data: A deep reinforcement learning approach. In *Proceedings of the International Conference on Multimodal Interaction*. 6–15.

[49] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6

(2017), 103–108.

[50] Henk WR Schreuder, Jan EU Persson, Richard GH Wolswijk, Ingmar Ihse, Marlies P Schijven, and René HM Verheijen. 2014. Validation of a novel virtual reality simulator for robotic surgery. *The Scientific World Journal* 2014 (2014).

[51] Chaopeng Shen. 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research* 54, 11 (2018), 8558–8593.

[52] Ajay Shrestha and Ausif Mahmood. 2019. Review of deep learning algorithms and architectures. *IEEE access* 7 (2019), 53040–53065.

[53] Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. 2019. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6163–6170.

[54] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Proceedings of the Conference on Robot Learning*. PMLR, 477–490.

[55] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[56] Daniel Szafir and Bilge Mutlu. 2012. Pay attention! Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 11–20.

[57] Yujie Tao and Pedro Lopes. 2022. Integrating Real-World Distractions into Virtual Reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[58] Nirmalya Thakur and Chia Y Han. 2021. An ambient intelligence-based human behavior monitoring framework for ubiquitous environments. *Information* 12, 2 (2021), 81.

[59] Sebastian Thrun. 2002. Probabilistic robotics. *Commun. ACM* 45, 3 (2002), 52–57.

[60] Zhuowen Tu and Song-Chun Zhu. 2002. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 5 (2002), 657–673.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[62] Luiz G Véras, Anna KF Gomes, Guilherme AR Dominguez, and Alexandre T Oliveira. 2022. Drivers' attention detection: a systematic literature review. *arXiv preprint arXiv:2204.03741* (2022).

[63] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. 2020. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters* 5, 2 (2020), 713–720.

[64] Deborah Rohm Young, Marie-France Hivert, Sofiya Alhassan, Sarah M Camhi, Jane F Ferguson, Peter T Katzmarzyk, Cora E Lewis, Neville Owen, Cynthia K Perry, Juned Siddique, et al. 2016. Sedentary behavior and cardiovascular morbidity and mortality: a science advisory from the American Heart Association. *Circulation* 134, 13 (2016), e262–e279.

[65] Zhenliang Zhang. 2021. Symmetrical cognition between physical humans and virtual agents. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 587–588.

[66] Zhenliang Zhang. 2023. Building Symmetrical Reality Systems for Cooperative Manipulation. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 751–752.

[67] Zhenliang Zhang, Cong Wang, Dongdong Weng, Yue Liu, and Yongtian Wang. 2019. Symmetrical reality: Toward a unified framework for physical and virtual reality. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 1275–1276.

[68] Zhenliang Zhang, Dongdong Weng, Haiyan Jiang, Yue Liu, and Yongtian Wang. 2018. Inverse augmented reality: A virtual agent's perspective. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 154–157.

[69] Nan Zhao, Asaph Azaria, and Joseph A Paradiso. 2017. Mediated atmospheres: A multimodal mediated work environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–23.

[70] Song-Chun Zhu, David Mumford, et al. 2007. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2, 4 (2007), 259–362.