Exploring the Effects of Self-Correction Behavior of an Intelligent Virtual Character during a Jigsaw Puzzle Co-Solving Task

MINSOO CHOI and SIQI GUO, Purdue University, West Lafayette, IN, USA
ALEXANDROS KOILIAS, University of the Aegean, Mytilene, Greece and University of the Peloponnese, Nafplio, Greece
MATIAS VOLONTE, Clemson University, Clemson, SC, USA
DOMINIC KAO and CHRISTOS MOUSAS, Purdue University, West Lafayette, IN, USA

Although researchers have explored how humans perceive the intelligence of virtual characters, few studies have focused on the ability of intelligent virtual characters to fix their mistakes. Thus, we explored the self-correction behavior of a virtual character with different intelligence capabilities in a within-group design ($N = 23$) study. For this study, we developed a virtual character that can solve a jigsaw puzzle whose self-correction behavior is controlled by two parameters, namely, *Intelligence* and *Accuracy of Self-correction.* Then, we integrated the virtual character into our virtual reality experience and asked participants to co-solve a jigsaw puzzle. During the study, our participants were exposed to five experimental conditions resulting from combinations of the *Intelligence* and *Accuracy of Self-correction* parameters. In each condition, we asked our participants to respond to a survey examining their perceptions of the virtual character's intelligence and awareness (private, public, and surroundings awareness) and user experiences, including trust, enjoyment, performance, frustration, and desire for future interaction. We also collected application logs, including participants' dwell gaze data, completion times, and the number of puzzle pieces they placed to co-solve the jigsaw puzzle. The results of all the survey ratings and the completion time were statistically significant. Our results indicated that higher levels of *Intelligence* and *Accuracy of Self-correction* enhanced not only our participants' perceptions of the virtual character's intelligence, awareness (private, public, and surroundings), trustworthiness, and performance but also increased their enjoyment and desire for future interaction with the virtual character while reducing their frustration and completion time. Moreover, we found that as the *Intelligence* and *Accuracy of Self-correction* increased, participants had to place fewer puzzle pieces and needed less time to complete the jigsaw puzzle. Finally, regardless of the experimental condition to which we exposed our participants, they gazed at the virtual character for more time compared to the puzzle pieces and puzzle goal in the virtual environment.

CCS Concepts: • **Computing methodologies** → **Virtual reality**; **Intelligent agents**; • **Human-centered computing** → **User studies**;

Authors' Contact Information: Minsoo Choi, Purdue University, West Lafayette, IN, USA; e-mail: choi714@purdue.edu; Siqi Guo, Purdue University, West Lafayette, Indiana, USA; e-mail: guo477@purdue.edu; Alexandros Koilias, University of the Aegean, Mytilene, Greece and University of the Peloponnese, Nafplio, Greece; e-mail: alex.kilias@go.uop.gr; Matias Volonte, Clemson University, Clemson, SC, USA; e-mail: mvolont@clemson.edu; Dominic Kao, Purdue University, West Lafayette, IN, USA; e-mail: kaod@purdue.edu; Christos Mousas (Corresponding author), Purdue University, West Lafayette, IN, USA; e-mail: cmousas@purdue.edu.

ACM 2160-6463/2024/9-ART23
https://doi.org/10.1145/3688006

Additional Key Words and Phrases: virtual reality, virtual character, intelligence, self-correction, jigsaw puzzle, puzzle co-solving

ACM Reference format:
Minsoo Choi, Siqi Guo, Alexandros Koilias, Matias Volonte, Dominic Kao, and Christos Mousas. 2024. Exploring the Effects of Self-Correction Behavior of an Intelligent Virtual Character during a Jigsaw Puzzle Co-Solving Task. *ACM Trans. Interact. Intell. Syst.* 14, 3, Article 23 (September 2024), 33 pages.
https://doi.org/10.1145/3688006

1 Introduction

Due to the democratization of **virtual reality (VR)** technologies, VR applications have gained unprecedented popularity [68]. VR has been successfully used in domains including training [52, 98], education [1], and games [51, 53]. In these domains, several applications allow VR users to interact with virtual characters [45, 47, 63–65, 92]. Specifically, virtual characters can observe and interact with virtual environments [100] and react to pre-defined events [20, 54, 69]. They can also "understand" the context of situations [84] and collaborate with others to achieve goals [5]. Furthermore, verbal or non-verbal communications [101] and emotional expressions [77, 92] are simulated to enhance the realism of interactions with virtual characters. These abilities enable virtual characters to deliver knowledge or information [16, 76] and collaborate with people to complete specific tasks [35, 51].

As virtual characters become more sophisticated, several researchers have explored how humans interact and experience them. However, although virtual characters can be scripted to perform a given task efficiently (i.e., act optimally and behave intelligently), exploring how human-like characteristics and behaviors can impact human perceptions of virtual characters is also important. For example, understanding how humans perceive mistakes made by virtual characters can help us further develop human-like intelligent virtual characters [36]. However, although some work has been conducted to understand how humans perceive agents or robots when the latter make grammatical mistakes [74, 87], malfunction [57], and perform unexpected movements [30], less attention has been given to how the self-correction behavior of a virtual character (i.e., its ability to fix its own mistake) could impact human perceptions of that virtual character.

To explore human perceptions of the self-correction behavior of a virtual character, we implemented a VR application in which a user and a virtual character collaborate to solve a jigsaw puzzle, which has been considered a cognitively demanding task identified by previously conducted research [29, 39]. The self-correction behavior we implemented enabled the virtual character to become aware that a wrong action had been performed and "self-correct" that action (i.e., the virtual character was scripted to pick up a wrongly placed puzzle piece and place it in a new [either right or wrong] position on the puzzle board). To control the behavior of our virtual character, we used two parameters, namely, *Intelligence* and *Accuracy of Self-correction*. The *Intelligence* parameter denotes the probability of placing the puzzle piece in the correct spot. For example, the virtual character always places the puzzle pieces correctly if we set *Intelligence* to 100%. The other parameter, *Accuracy of Self-correction*, is the probability that the virtual character will fix its mistake correctly. When the virtual character has 0% *Accuracy of Self-correction*, it always picks up its last wrongly placed puzzle piece but again places it in the wrong spot. In contrast, when the virtual character has 100% *Accuracy of Self-correction*, it always places its last wrongly placed puzzle piece in the correct spot.

In this project, we developed a virtual character programmed to assist study participants in co-solving a jigsaw puzzle, with its actions guided by a user-defined probability of *Intelligence* and

Accuracy of Self-correction when placing the jigsaw puzzle pieces. Although the behavior of the virtual character might not entirely align with conventional definitions of intelligence, we propose that it can still be regarded as "intelligent." According to Fissler et al. [29], solving jigsaw puzzles involves various cognitive skills, including visual perception for recognizing shapes and patterns, constructional praxis for coordinating visual and motor information, mental rotation for aligning pieces, and cognitive flexibility for adjusting strategies. Additionally, these tasks require cognitive speed, perceptual reasoning for developing strategies, and working and episodic memory to keep track of the associations between puzzle pieces. Given these demands, we argue that the virtual character's ability to address these cognitive challenges supports its classification as "intelligent."

To explore how study participants perceived the virtual character when they were co-solving the jigsaw puzzle, we conducted a within-group study ($N = 23$) and asked our participants to collaborate with the virtual character and co-solve the jigsaw puzzle in each of the five experimental conditions we developed. Our experimental conditions were the results of combining *Intelligence* and *Accuracy of Self-correction* parameters: (1) 0% Intelligence without Self-correction, (2) 0% Intelligence and 0% Accurate Self-correction, (3) 0% Intelligence and 50% Accurate Self-correction, (4) 0% Intelligence and 100% Accurate Self-correction, and (5) 100% Intelligence without Self-correction. For example, in the 0% Intelligence with 100% Accurate Self-correction condition, the virtual character places the puzzle piece in the wrong spot on the board, then picks it up again and places it in the correct spot. After finishing each condition, we asked participants to self-report their experience with the virtual character by answering a survey. The survey comprised questions examining 10 variables: perceived intelligence, intelligence comparison (participants compared their intelligence against the virtual character's intelligence), virtual character's private awareness, virtual character's public awareness, virtual character's surroundings awareness, trust, performance, enjoyment, frustration, and desire for future interaction as well as an open-ended question for our participants to provide additional feedback. We also collected application logs, including the participants' dwell gazing (virtual character, puzzle pieces, and puzzle goal), completion time, and the number of puzzle pieces participants placed during co-solving the jigsaw puzzle.

We organized this article as follows. In Section 2, we discuss work related to our project. In Section 3, we present details of our implementation and methodology. In Section 4, we present our results, which are discussed in Section 5, along with our study's limitations. Finally, in Section 6, we draw conclusions and discuss potential future work.

2 Related Work

2.1 Human–Agent Interaction

Researchers in human–computer interaction have extensively studied how humans interact with computer systems [38]. Human–agent interaction is an extension of human–computer interaction as it regards agents as interactive systems. Numerous researchers have defined the concept of agents. Norman [72] described them as "...forth images of human-like automatons, working without supervision on tasks thought to be for our benefit, but not necessarily to our liking." Lewis [50] also focused on automation as the concept of agents in human–computer interaction. Based on these concepts, various agents, such as virtual agents or robots, exist within the human–agent interaction research area.

Several researchers have investigated human–agent interaction using robots. Bradshaw et al. [10] referred to the derivation of robots to describe agents. Burghart et al. [11] proposed an approach to train an anthropomorphic robot to solve a jigsaw puzzle like a child with a tutor. They recorded a video of a child solving a jigsaw puzzle with the tutor and converted it to an applicable format to train the robot. In the proposed approach, the trained robot solved the jigsaw puzzle cooperatively based on human instruction or guidance. Giuliani and Knoll [32] assigned instructive or supportive roles to robots in human–robot interaction to understand how people react. They found that people did not prefer the robot to take either role and behaved as the counterpart of whichever role it played.

Regarding virtual agents, Cerekovic et al. [14] integrated virtual characters to let participants interact with them and analyzed the interaction through the perspective of personality traits and non-verbal cues. They built regression models to predict interaction experiences and found that the best predictions could be made using both personality traits and non-verbal cues. Morton and Jack [62] proposed a computer-assisted language learning program with a virtual agent that spoke with users and gave feedback regarding grammatical and ungrammatical utterances. The virtual agent reacted to the user's speech and changed its dialog based on the communication difficulties to support the user's language learning. Furthermore, Cavazza et al. [13] proposed a prototype for interactive storytelling based on user intervention. In their prototype, the user could interact with the virtual agents, and this interaction affected the agents' subsequent behaviors and triggered changes in the storyline.

2.2 Collaboration with Virtual Agents

The Merriam-Webster dictionary defines collaboration[1] as "working jointly with others or together, especially in an intellectual endeavor." Rickel and Johnson [78] used the term "collaboration" to describe agents helping users learn given tasks and thus meet their objectives. Researchers have explored how people collaborate with virtual agents. Andrist et al. [2] implemented a virtual agent that taught users the task of sandwich-making to improve the user experience in human–agent interaction through producing and detecting gaze cues. The authors found that the bidirectional gaze of the virtual character (producing its gaze and responding to the user's gaze) positively impacted the number of errors, and the virtual character's response to the participant's gaze improved the degree of coordination during collaboration.

Collaborative virtual agents have also been applied to games. Merritt et al. [60] integrated an artificial agent into a cooperative game to compare the users' perception of the risk-taking action of different types of teammates, and the artificial agent collaborated with players to win the game through risk-taking action. The authors found that the players noticed more risk-taking actions when they thought about collaborating with humans than with artificial intelligence. Daronnat [22] explored the human–agent trust relationship in collaborative games through different aspects of agents, such as predictability or type of errors. The author stated that an error caused by the agent's inaction impacted trust and performance less negatively than planning or commission errors.

While the previously mentioned studies focused on human–agent interaction, other studies have considered agent–agent interaction. Liu et al. [51] implemented collaboration between virtual agents driven by behavior trees to measure the degree of collaboration in gameplay. Later, when Liu et al. evaluated their game levels, they found a strong correlation between the degree of collaboration data provided by virtual agents and their study participants. Cavazza et al. [13] described the interaction between agents as an intervention. Specifically, as in a user intervention, one agent could interfere with other agents and affect the storytelling. Baker et al. [5] presented a reinforcement learning algorithm to train virtual agents to play hide and seek cooperatively. They indicated that collaborations between virtual agents were observed as a strategy for winning in team play.

[1] https://www.merriam-webster.com/dictionary/collaboration

2.3 Awareness in Human–Agent Interaction

Some studies on human–agent interaction have found that some virtual agents can detect the surrounding environment, such as the progress of a task [16], and decide their subsequent behavior. These dynamically allocated behaviors make the virtual agents look more aware. According to the Merriam-Webster dictionary, awareness[2] is "the quality or state of being aware: knowledge and understanding that something is happening or exists," and it has various targets.

Munir et al. [67] defined situational awareness as "the perception of entities in the environment, comprehension of their meaning, and projection of their status in the near future." Similarly, Livnat et al. [55] indicated that situational awareness is "the ability to identify, process, and comprehend the critical elements of information about what is happening." Situational awareness can be applied to decision-making [55], such as remotely controlling urban search and rescue robots [99]. Govern and Marsch [33] proposed the situational self-awareness scale, which comprises three types of awareness: private, public, and surroundings. The authors stated that private awareness is related to attention to one's inner feelings, public awareness is based on attention to how one shows oneself to others, and surroundings awareness describes attention to the environment. Ijaz et al. [41] defined three types of awareness (environment awareness, self-awareness, and interaction awareness) and provided detailed information for all three types to implement an aware virtual agent. The authors found that the aware virtual agent was more believable than the unaware one. Contrarily, the awareness target of virtual agents can be the users. McNeely-White et al. [58] presented a virtual agent that was aware of aspects of the users, such as gestures or gazes, through videos and depth sensors. They found that the shared perception and user awareness of the virtual character made the user feel and interact with it as if it were a person. Furthermore, Tan et al. [89] implemented a location-aware virtual character based on the locations of users and objects and stated that people perceived higher presence and adaptivity from the location-aware virtual character than a virtual character that was unaware of the location of users and objects.

2.4 Perceived Intelligence

Norman [72] used the term "intelligent" to describe "agents." Numerous studies demonstrate that intelligence is one of the main components of agents [5, 32, 59, 88, 90]. Perceived intelligence is humans' perceptions of agents' intelligence [66], and researchers have explored various factors that can affect it. Deshmukh et al. [24] focused on the relationship between human perception and the understandability of robot gestures. They found a correlation between perceived intelligence and understandability. Lee et al. [49] applied an anthropomorphic layer to a virtual agent, showing that the virtual agent's appearance positively affected perceived intelligence. They identified that anthropomorphism due to a human-like appearance caused higher social presence and positively impacted perceived intelligence. Choi et al. [17] explored the effect of virtual characters' appearance and voice mismatch on perceived intelligence and found that virtual characters with a robot-like appearance had a higher perceived intelligence than those that looked like humans. Finally, Bartneck et al. [6] explored the connection between a robot's perceived intelligence, animacy, and design and found a correlation between animacy and perceived intelligence.

2.5 Self-Correction

Even if an intelligent virtual character is designed to perform tasks correctly, it is hard to guarantee that a virtual character can treat all cases without errors. Thus, several researchers have studied how people perceive virtual characters when they make mistakes. According to Wang et al. [94], even if virtual characters make errors such as unresponsiveness, irrelevant responses, or other

[2] https://www.merriam-webster.com/dictionary/awareness

conversational mistakes, they can still provide social influence and impact user interaction. However, according to Lucas et al. [56], such errors affect human task performance and reduce a virtual character's persuasive ability. In contrast, according to Skarbez et al. [83], there is a strong correlation between error metrics and the perceived quality of interaction with the virtual character.

However, when a virtual character performs tasks incorrectly, it should be able to identify and correct its mistakes, resulting in a self-correction behavior. Satne [82] presented three components of self-correction: the application of concepts, the ability to evaluate the applications of concepts, and the modification of the application of concepts based on the evaluation. Self-correction behaviors vary in terms of how the corrective actions are applied. Lasecki and Bigham [48] researched how to leverage self-correction from crowds and proposed two types of self-correction: averaging and voting. They found that self-corrections helped crowds reach the appropriate correction before the final decision without the identification of invalid input. Ming et al. [61] proposed a framework that enables robot self-correction. They used a perception detector to collect environmental information and determine whether there were errors. Based on the decisions, the corrector assigned appropriate feedback, such as high-level or low-level feedback, to correct the error. The authors applied the framework to a robot in a real environment and showed its error detection and correction capacity. Such findings highlight the need to conduct studies to understand how mistakes made by virtual characters could impact human perceptions of and interactions with them.

2.6 Contributions

Researchers on human–agent interaction have extensively examined how virtual agents and robots collaborate with humans, focusing on factors like awareness [67, 89], perceived intelligence [6, 17, 49], and user experience [2, 22, 60]. However, there is a lack of research specifically investigating how self-correction behaviors in virtual characters impact these factors. Most studies have explored basic interaction dynamics and task performance [10, 32, 62], but the effects of self-correction accuracy and intelligence remain underexplored [48, 61]. Our research addresses this gap by examining how self-correction behavior influences perceived intelligence, awareness, user experience, and behavioral responses, providing a more comprehensive understanding of effective human–virtual character interactions.

Specifically, the contributions of our work are as follows. First, we introduce algorithms that control how a virtual character could solve a jigsaw puzzle based on *Intelligence* and *Accuracy of Self-correction* parameters. Second, we conducted a user study to understand further how the self-correction behavior of an intelligent virtual character could impact humans' perception of the virtual character and the user experience of the developed application and task with which study participants were asked to interact. Finally, we think our findings could help researchers explore further how self-correction behavior can be implemented in intelligent virtual characters, an underexplored research direction of the human-virtual character interaction field.

2.7 Research Questions

We identified several research questions for our study to understand how a virtual character's self-correction behavior could impact participants' perceptions of the virtual character as well as their experiences as users. Specifically, we examine four research questions:

- *Intelligence*: This research question explores how participants perceive the intelligence of a virtual character when it exhibits self-correction behaviors. It includes investigating subjective perceptions of intelligence and comparative ratings when we asked our participants to evaluate the virtual character's intelligence.

  - *RQ1*: How do the self-correction behaviors of a virtual character impact participants' perceptions and comparative ratings of that virtual character's intelligence?
- *Awareness*: This research question delves into how self-correction behaviors influence participants' views on the virtual character's awareness. It covers: (1) Private Awareness, the perception of the virtual character's self-awareness or internal state; (2) Public Awareness, the perception of the virtual character's awareness of others and social contexts; and (3) Surroundings Awareness, the perception of the virtual character's awareness of the physical environment and situational context.
  - *RQ2*: How do the self-correction behaviors of a virtual character impact participants' perceptions of the virtual character's awareness in various contexts?
- *User Experience*: This research question assesses the overall user experience when interacting with a self-correcting virtual character. Key aspects include: (1) Trust, the level of trust participants have in the virtual character; (2) Performance, how well participants perform tasks in conjunction with the virtual character; (3) Enjoyment, the degree of enjoyment participants experience during the interaction; (4) Frustration, the amount of frustration felt by participants; and (5) Desire for Future Interaction, participants' willingness to engage with the virtual character again in the future.
  - *RQ3*: How do the self-correction behaviors of a virtual character impact participants' user experience, including trust, performance, enjoyment, frustration, and willingness to interact in the future?
- *Behavioral Responses*: This research question focuses on the observable behavioral responses of participants. It includes: (1) Dwell Gazes, where participants focus their gaze, specifically on the virtual character, the puzzle goal, and the puzzle pieces; (2) Task Completion Times, how quickly participants complete tasks when interacting with the self-correcting virtual character; and (3) Number of Puzzle Pieces, how many puzzle pieces participants placed on the puzzle board to solve the jigsaw puzzle.
  - *RQ4*: How do the self-correction behaviors of a virtual character impact participants' behavioral responses, including gaze patterns, task completion times, and the number of puzzle pieces they place when co-solving the jigsaw puzzle?

3 Materials and Methods

3.1 Participants

We conducted an *a priori* power analysis using the G*Power version 3.1 software [27] to determine the appropriate sample size for our study. For an 80% power (1-β error probability), a small effect size of f = .25 [18], one group with five repeated measures, a non-sphericity correction ϵ = .90, and an α = .05, the analysis recommended a minimum of 22 participants. We recruited 23 participants (age: M = 23.82, SD = 4.26) through e-mails sent to our university's students and class announcements. All of our participants were undergraduate and graduate students at a Midwest U.S. university. Of the sample, 15 were males (age: M = 24.60, SD = 4.35), and 8 were females (age: M = 22.37, SD = 3.92). All participants had prior VR experience.

3.2 Implementation

We developed a VR jigsaw puzzle application in the Unity game engine version 2020.3.20. We used Meta's Quest 1 as a VR **head-mounted display (HMD)** and a Dell Alienware Aurora R7 desktop computer with Intel Core i7, NVIDIA GeForce RTX 2080, and 32GB RAM for the implementation of our application and study. Our VR application comprises a virtual environment (see Figure 1), an intelligent virtual character (see Figure 2), a dialog manager, and user interaction tools.

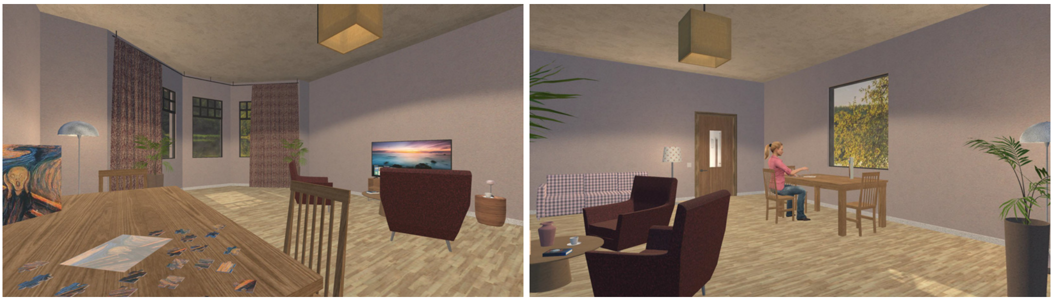

Fig. 1. We designed a semi-realistic living room as the virtual environment where we immersed the participants in our study.

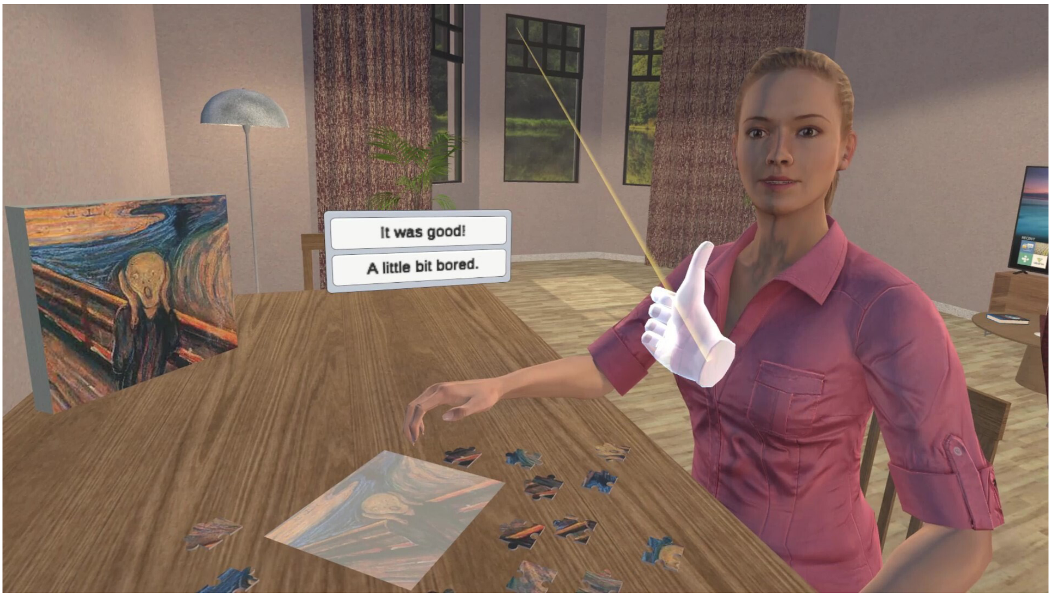

Fig. 2. We applied an L-shaped formation to support social interaction between the participant and the virtual character. The virtual character is on the participant's right side.

3.2.1 Virtual Environment. The virtual environment of our application was a semi-realistic living room three-dimensional model (see Figure 1). In the living room, we added furniture and appliances to provide a cozy atmosphere for the participants. Both the virtual character and participant sat on chairs around the table. We applied an L-shaped formation (see Figure 2) from the F-formations models to support social interaction between the participant and the virtual character [75]. Thus, the participants could see the virtual character sitting to their right. We want to note that in our study, we used a female virtual character in all conditions to standardize the stimulus across all participants.

We placed all the puzzle pieces, the puzzle board, and the puzzle targets on the table. During our application's development and testing process, we conducted a preliminary study with our laboratory members to explore the application's flow, identify bugs, and determine the optimal number and size of puzzle pieces to eliminate any negative effects on participants' experiences during the study. We realized that fewer pieces (25 in our case) would make the jigsaw puzzle

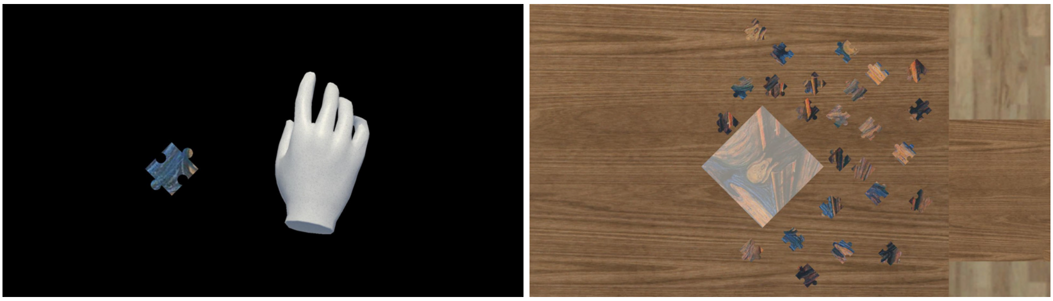

Fig. 3. Left: The size of a puzzle piece. Right: All (25 total) puzzle pieces and the semi-transparent puzzle board.

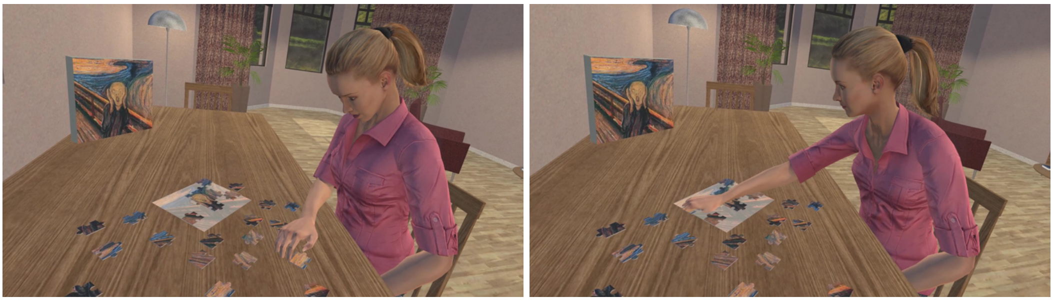

Fig. 4. The virtual character picks a puzzle piece and places it in a spot on the puzzle board. The brain system decides which puzzle piece the virtual character picks up and where the virtual character places it.

co-solving process more efficient, while more pieces would frustrate participants. In total, our puzzle was composed of 25 puzzle pieces, each one 4 × 4 cm in size. If we had too many puzzle pieces, this would significantly increase the duration of the experiment and might cause fatigue and loss of motivation among participants. We used a semi-transparent puzzle board to help the participants find the appropriate spot to place each puzzle piece, and the initial distributions of puzzle pieces remained consistent across all conditions (see Figure 3).

3.2.2 The Intelligent Virtual Character. Our intelligent virtual character can co-solve the puzzle with the participant. We implemented and assigned brain and animation systems to make our virtual character capable of solving the jigsaw puzzle (see Figure 4) and correcting its mistakes (see Figure 5). We want to note that the virtual character was not scripted to correct potential mistakes made by participants; however, the participants were able to fix the mistakes made by the virtual character. For the brain system, we integrated the *Intelligence* and *Accuracy of Self-correction* parameters. The brain system decides the state and behavior of the virtual character, and the animation system animates the virtual character according to the decision of the brain system. We provide a video as supplementary material that demonstrates the behaviors of our virtual character.

Brain System. The brain system (see Algorithm 1) controls how our virtual character solves the puzzles based on the user-defined *Intelligence* and *Accuracy of Self-correction* parameters. These parameters allow the virtual character to pick up a puzzle piece, place it in the right or wrong target spot, and trigger the self-correction behavior if required. We want to note that we did not implement a turn-taking mechanism for our virtual character to solve the jigsaw puzzle, as jigsaw

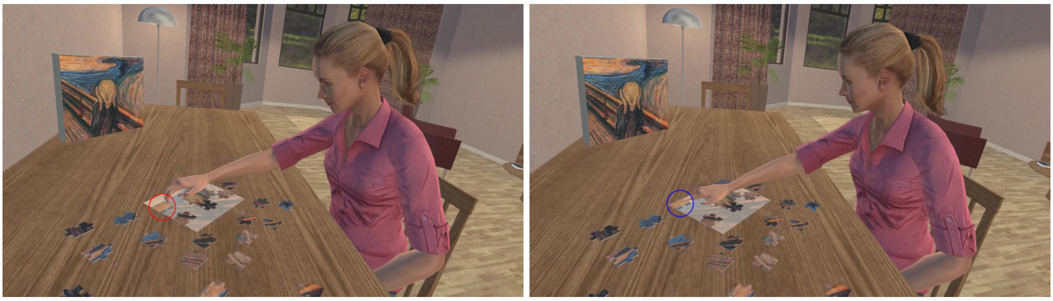

Fig. 5. An example of the self-correction behavior. The virtual character picks the last interacted puzzle piece and places it in the right spot. The red circle (left) shows the wrong puzzle piece, and the blue circle (right) shows the corrected one.

puzzles are not considered turn-based games like chess or backgammon. Thus, the virtual character did not wait for the participant to place puzzle pieces.

The brain system has eight inputs: R, U, P, I, B, A, V, V_l, and S. R is a list of puzzle targets not yet solved, U is a list of puzzle pieces that can be picked up, P is a list of pairs of puzzle pieces and their answers, I is the virtual character's *Intelligence* with a range from 0% to 100%, B is a Boolean value to indicate the availability of self-correction, A is *Accuracy of Self-correction* with a range from 0% to 100%, V is the current puzzle piece interacted with by the virtual character, V_l is the last puzzle piece interacted with by the virtual character, and S is the virtual character's current state. Specifically, the brain system has a set of pre-defined S: *PickUp* (see Algorithm 2), *Place* (see Algorithm 3), $SelfCorrectionPickUp$ (see Algorithm 4), $SelfCorrectionPlace$ (see Algorithm 5), and $Wait$.

Behavior Functions. Each state, except the $Wait$, includes an assigned function that decides how the virtual character behaves with the puzzle piece with which it currently interacts and updates the virtual character's state and that puzzle piece. The brain system has four behavior functions: `PickUp`, `Place`, `SelfCorrPickUp`, and `SelfCorrPlace`. The `PickUp` function belongs to the $PickUp$ state, chooses a puzzle piece from U, and makes the virtual character pick it up. However, if there is no available puzzle piece, the state goes to the $Wait$ state, and the virtual character waits until there is at least one available puzzle piece. The `Place` function belongs to the *Place* state and lets the virtual character place the puzzle piece in a specific spot. Specifically, if the brain system does not allow the virtual character to perform its self-correction, the target spot is decided by I. Otherwise, the virtual character will place the puzzle piece incorrectly. Moreover, the `Place` function decides the following states according to B. If B is true, the next state will be the $SelfCorrectionPickUp$ state. Otherwise, it will be the $PickUp$ state. The `SelfCorrPickUp` function belongs to the $SelfCorrectionPickUp$ state. It has a fixed input, namely, the last puzzle piece interacted with, to allow the virtual character to pick up the last puzzle piece interacted with for the self-correction behavior. It also has a fixed update of the state, the $SelfCorrectionPlace$ state, to complete the self-correction behavior. Finally, the `SelfCorrPlace` function belongs to the $SelfCorrectionPlace$ state. It chooses the spot where the puzzle price should be placed by A instead of I, and the virtual character places the puzzle piece to fix its previous mistake. It updates the state to the $PickUp$ state to let the virtual character solve the puzzle continuously. Note that the local variable T is the target spot where the puzzle piece should be placed, D is a random variable between 0% and 100% to determine the behavior based on *Intelligence* or *Accuracy of Self-correction*, and V^a is the correct spot of V mapped by P.

Algorithm 1: Brain System State Decision Algorithm

```
Input:
$R \in \{R_1, \cdots, R_l\}$                          ▷ $R$ is a list of puzzle targets not yet solved
$U \in \{U_1, \cdots, U_m\}$                          ▷ $U$ is a list of puzzle pieces that can be picked up
$P \in \{(x_1, x_1^a), \cdots, (x_n, x_n^a)\}$        ▷ $P$ is a list of pairs of puzzle pieces and their answers
$I$                                                   ▷ $I$ is the virtual character Intelligence (0% - 100%)
$B$,                                                  ▷ $B$ is a Boolean value to indicate the availability of self-correction
$A$,                                                  ▷ $A$ is Accuracy of Self-correction (0% - 100%)
$V$,                                                  ▷ $V$ is the current puzzle piece interacted with by the virtual character
$V_l$,                                                ▷ $V_l$ is the last puzzle piece interacted with by the virtual character
$S$,                                                  ▷ $S$ is the virtual character's current state
Output:
$V_u$                                                 ▷ $V_u$ is the updated current puzzle piece interacted with by the virtual character
$V_{lu}$                                              ▷ $V_{lu}$ is the updated last puzzle piece interacted with by the virtual character
$S_u$                                                 ▷ $S_u$ is the virtual character's updated state
function BrainSystem($R$, $U$, $P$, $I$, $B$, $A$, $V$, $V_l$, $S$)
    switch $S$ do
        case $PickUp$
            $V_u, S_u \leftarrow$ PickUp($U$)
        case $Place$
            $V_u, V_{lu}, S_u \leftarrow$ Place($R$, $P$, $I$, $V$, $B$)
        case $SelfCorrectionPickUp$
            $V_u, S_u \leftarrow$ SelfCorrPickUp($V_l$)
        case $SelfCorrectionPlace$
            $V_u, S_u \leftarrow$ SelfCorrPlace($R$, $P$, $A$, $V$)
        case $Wait$
            if $U > 0$ then
                $S_u \leftarrow PickUp$
            end if
    return $V_u, S_u$
end function
```

Animation System. As the brain system decides the state of the virtual character, the animation system controls the latter's movement. We implemented the full-body forward and backward inverse kinematic solver [3] to allow the virtual character to perform picking and placing tasks. During the puzzle-solving process, the system is given the chosen puzzle piece or target spot as input, which drives the end-effector (the virtual character's right arm) to reach the target spot. Because the end-effector is connected with the upper body of the virtual character, the inverse kinematics solver controls and animates the virtual character's upper body parts (i.e., shoulders and spine).

We also implemented gaze targets. Specifically, we scripted the virtual character to coordinate its gaze with its right hand while trying to pick up and place the chosen puzzle piece (it moves its head to gaze at its right hand when performing the pick-it-up animation). In addition, we implemented

Algorithm 2: Virtual Character Pick Up Puzzle Algorithm

```
Input:
$U \in \{U_1, \cdots, U_m\}$          ▷ $U$ is a list of puzzle pieces that can be picked up
Output:
$V_u$          ▷ $V_u$ is the updated current puzzle piece interacted with by the virtual character
$S_u$          ▷ $S_u$ is the virtual character's updated state
function PickUp($U$)
    if $U > 0$ then
        Choose $T$ from $U$ Randomly
        Pick up $T$
        $V_u \leftarrow T$
        $S_u \leftarrow Place$
    else
        $S_u \leftarrow Wait$
    end if
    return $V_u$, $S_u$
end function
```

eye-blink animation and assigned an idle motion with a sitting pose to make our virtual character's movements look more realistic.

3.2.3 Dialog Manager. We implemented a conversational virtual character [93] controlled by a dialog manager to provide pre-defined dialogs in our VR jigsaw puzzle application. It provides dialogs in three phases: the beginning, middle, and end of the VR experience. Each dialog included a set of pre-defined answers. The dialog manager detects the progress of solving the puzzle by detecting the number of unsolved puzzle pieces. More specifically, the first dialog phase is generated when all puzzle pieces are unsolved, the second dialog phase is generated when half are unsolved, and the last phase is generated when there are no unsolved puzzle pieces. We used Microsoft's Azure[3] text-to-speech service to generate the dialogs and the SALSA LipSync Suite[4] from Unity Asset Store to synthesize the lip-sync animation. Additionally, we assigned humming to the virtual character in randomly chosen timesteps to make participants think that the virtual character was thinking about its decisions. We did so because prior studies have shown that dialogs provided more engaged experiences [28], trust, and rapport [8, 42].

3.2.4 User Interaction Tool. We used the Oculus Integration Toolkit to support user interaction in the VR jigsaw puzzle experience. It provided simulated hand models based on the input signals from controllers, and the simulated hands helped the user grab the puzzle piece and place it on the puzzle board or table through natural gestures. The toolkit also supported **user interface (UI)** interaction based on ray casting, so users could point to UI components directly and interact with them, such as clicking the button or moving the slide. In our VR puzzle co-solving experience, we used the UI to let participants choose and answer from the implemented dialogues.

[3] https://azure.microsoft.com/en-us/products/cognitive-services/text-to-speech
[4] https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442

Algorithm 3: Virtual Character Place Puzzle Algorithm

```
Input:
$R \in \{R_1, \cdots, R_l\}$        ▹ $R$ is a list of puzzle targets not yet solved
$P \in \{(x_1, x_1^a), \cdots, (x_n, x_n^a)\}$        ▹ $P$ is a list of pairs of puzzle pieces and their answers
$I$        ▹ $I$ is the virtual character Intelligence (0% - 100%)
$V$        ▹ $V$ is the current puzzle piece interacted with by the virtual character
$B$        ▹ $B$ is a Boolean value to indicate the availability of self-correction
Output:
$V_u$        ▹ $V_u$ is the updated current puzzle piece interacted with by the virtual character
$V_{lu}$        ▹ $V_{lu}$ is the updated last puzzle piece interacted with by the virtual character
$S_u$        ▹ $S_u$ is the virtual character's updated state
function PLACE($R$, $P$, $I$, $V$, $B$)
    if $B$ then
        while $T = V^a$ do
            Choose $T$ from $R$ Randomly
        end while
        $S_u \leftarrow SelfCorrectionPickUp$
    else
        Choose $D$ from 0% to 100% Randomly
        if $D \leq I$ then
            Choose $T$ as $V^a$ from $P$ by using $V$
        else
            while $T = V^a$ do
                Choose $T$ from $R$ Randomly
            end while
        end if
        $S_u \leftarrow PickUp$
    end if
    Place $V$ on the $T$
    $V_{lu} \leftarrow V_u$
    $V_u \leftarrow NULL$
    return $V_u$, $V_{lu}$, $S_u$
end function
```

3.3 Experimental Conditions

We designed five experimental conditions to explore how self-correction behaviors controlled by *Intelligence* and *Accuracy of Self-correction* parameters affect users' perceptions and experiences. For our study, we used a within-group design to let participants make direct comparisons across the five experimental conditions. We examined the following conditions:

- ***0% Intelligence without Self-correction (LI).*** In this condition, we assigned a 0% probability of solving the puzzle and disabled self-correction behavior. The virtual character always places a puzzle piece in the wrong spot on the puzzle board and does not correct it later, thus not contributing to solving the puzzle.

Algorithm 4: Virtual Character Self Correction Pick Up Puzzle Algorithm

```
Input:
$V_l$        ▷ $V_l$ is the last puzzle piece interacted with by the virtual character
Output:
$V_u$        ▷ $V_u$ is the updated current puzzle piece interacted with by the virtual character
$S_u$        ▷ $S_u$ is the virtual character's updated state
function SelfCorrPickUp($V_l$)
    Choose $T$ from $V_l$
    Pick up $T$
    $V_u \leftarrow T$
    $S_u \leftarrow SelfCorrectionPlace$
    return $V_u$, $S_u$
end function
```

Algorithm 5: Virtual Character Self Correction Place Puzzle Algorithm

```
Input:
$R \in \{R_1, \cdots, R_l\}$        ▷ $R$ is a list of puzzle targets not yet solved
$P \in \{(x_1, x_1^a), \cdots, (x_n, x_n^a)\}$        ▷ $P$ is a list of pairs of puzzle pieces and their answers
$A$        ▷ $A$ is Accuracy of Self-correction (0% - 100%)
$V$        ▷ $V$ is the current puzzle piece interacted with by the virtual character
Output:
$V_u$        ▷ $V_u$ is the updated current puzzle piece interacted with by the virtual character
$S_u$        ▷ $S_u$ is the virtual character's updated state
function SelfCorrPlace($R$, $P$, $A$, $V$)
    Choose $D$ from 0% to 100% Randomly
    if $D \leq A$ then
        Choose $T$ as $V^a$ from $P$ by using $V$
    else
        while $T = V^a$ do
            Choose $T$ from $R$ Randomly
        end while
    end if
    Place $V$ on the $T$
    $V_u \leftarrow NULL$
    $S_u \leftarrow PickUp$
    return $V_u$, $S_u$
end function
```

- ***0% Intelligence with 0% Accurate Self-correction (LSC).*** In this condition, we assigned a 0% probability of solving the puzzle and a 0% probability of fixing the previous error. Hence, the virtual character always places a puzzle piece in the wrong spot on the puzzle board; then, the virtual character tries to correct the previous mistake by picking it up from the puzzle board, but it again places it in the wrong spot. As before, the virtual character does not contribute to solving the puzzle at all.

- ***0% Intelligence with 50% Accurate Self-correction (MSC).*** In this condition, we assigned a 0% probability of solving the puzzle and a 50% probability of fixing the previous error. Hence, the virtual character always places a puzzle piece in the wrong spot on the puzzle board; then, the virtual character tries to correct its previous mistake, but half of the time, it places the puzzle piece in the wrong spot (the other half of the fixes are correct). In this condition, the virtual character can contribute up to 50% to solving the puzzle.
- ***0% Intelligence with 100% Accurate Self-correction (HSC).*** In this condition, we assigned a 0% probability of solving the puzzle and a 100% probability of fixing the previous mistake. Hence, the virtual character always places a puzzle piece in the wrong spot on the puzzle board; then, the virtual character tries to correct its previous mistake and always places the puzzle piece in the right spot. In this condition, the virtual character can contribute up to 100% to solving the puzzle.
- ***100% Intelligence without Self-correction (HI).*** In this condition, we assigned a 100% probability of solving the puzzle and disabled self-correction behavior. Hence, the virtual character always places a puzzle piece on the right spot of the puzzle board, and it does not self-correct. In this condition, the virtual character can contribute up to 100% to solving the puzzle.

Although we could have had other conditions with different combinations of *Intelligence* and *Accuracy of Self-correction*, we limited the number of conditions in case the participants lost interest and thought the VR experience was tedious. We want to note that our conditions were inspired by Sarkar et al. [81], who implemented and explored interaction with three conditions of a faulty robot and one condition of a non-faulty robot. However, we extended the schema by Sarkar et al by implementing conditions that cover the continuum between a faulty (0% Intelligence without Self-correction) and non-faulty (100% Intelligence without Self-correction) virtual character while also encountering self-correction behavior. Finally, we would like to mention that we used Latin squares [95] to balance the conditions and eliminate first-order carry-over (residual) effects.

3.4 Ratings and Measurement

We collected questionnaire responses as subjective data and application logs as objective data to understand how a virtual character's self-correction behavior affects users' perceptions and experiences.

3.4.1 Survey. We developed a survey to understand how the self-correction behavior of a virtual character affects users' perceptions and experiences. The survey comprised 21 items that examined 10 variables: perceived intelligence, intelligence comparison, virtual character's awareness (private awareness, public awareness, and surroundings awareness), trust, performance, enjoyment, frustration, and desire for future interaction. The items for the perceived intelligence were taken from Moussawi and Koufaris [66], and we used them to understand how our study participants perceived the intelligence of the virtual character through the different conditions we implemented. The awareness scales (private awareness, public awareness, and surroundings awareness) were taken from Govern and Marsch [33] and were used to understand if the virtual character is aware of the mistakes, it would make our participants rate the virtual character's awareness higher. We adopted the items of the trust scale from the System Trust Scale developed by Jian et al. [43]. We developed all the other items (intelligence comparison, performance, enjoyment, frustration, and desire for future interaction) ourselves. We used a 7-point Likert scale for the questionnaire responses. We provided the questionnaire after each condition and asked participants to give feedback about their experience when the experiment was finished. We distributed the survey and feedback form using the Qualtrics online survey tool. We provide the survey we developed for our study in Table A1 in Appendix A.

3.4.2 Application Logs. We collected data from our VR jigsaw puzzle application to understand how participants interact with the virtual character. Specifically, we collected:

- *Virtual Character Dwell Gazing.* We measured (normalized time) how long a participant gazed at the virtual character's upper body (including face, arms, and torso) while solving the puzzle.
- *Puzzle Goal Dwell Gazing.* We measured (normalized time) how long a participant gazed at the puzzle goal while solving the jigsaw puzzle.
- *Puzzle Pieces Dwell Gazing.* We measured (normalized time) how long a participant gazed at the puzzle pieces while solving the jigsaw puzzle.
- *Completion Time.* We measured (in seconds) how much time our participants needed to co-solve the jigsaw puzzle with the virtual character.
- *Number of Puzzle Pieces.* We counted the number of puzzle pieces our participants placed on the puzzle board to co-solve the jigsaw puzzle with the virtual character.

We assessed participants' visual attention by projecting a ray from the position of the HMD in the direction of their view into the virtual environment. If the projected ray intersected with a geometry model in the environment, this information was recorded for subsequent analysis. This approach to determining visual attention underwent scrutiny before the main experiment. We conducted a preliminary study with two laboratory members, during which they consistently focused on objects. Our method was able to detect their gazes accurately. Additionally, we want to note that researchers have documented the successful implementation of categorizing visual interest through analysis of HMD position and viewpoint in peer-reviewed publications [12, 40, 92]. In the study, the duration (in milliseconds) of this collision was returned when the ray collided with an object or the virtual character. After the participant had solved the puzzle, the visual attention method returned measured time with a name tag so we could track participants' perspectives and how they interacted.

3.5 Procedure

When a participant arrived at our research laboratory for this study, the research team provided the consent form with key information about the experiment procedure. After participants signed the consent form, they proceeded to the next part of our study. Our university's Institutional Review Board approved our study and consent form. After completing the demographics questionnaire, participants put on the VR HMD and started the tutorial scene. The tutorial aimed to familiarize participants with grabbing and placing puzzle pieces in our VR puzzle game. We implemented this tutorial as a prior study showed that VR tutorials improve study participants' user experiences and performances [44]. The tutorial scene took place in the same virtual environment, but there was no virtual character, and the puzzle pieces were different than those used in the main study. Instead, there were four puzzle pieces and an instruction window. The tutorial provided two tasks through the instruction window: picking up and placing the puzzle pieces in the right spot and fixing wrongly placed puzzle pieces.

When the participant had completed the tutorial, the research team ran the VR puzzle application with a specified sequence based on the Latin squares [95] ordering method. While the participant was solving the puzzle, the research team provided no information, such as whether the virtual character would fix its mistake. We also did not provide specific guidelines to our participants on how to complete the task (e.g., to complete it as soon as possible). Once the participants had completed the given condition, the research team asked them to take off the VR HMD and answer the questionnaire on Qualtrics on a desktop computer. At the end of each condition, the research team asked whether the participant wanted an additional break before starting the following condition. This process was repeated for each condition.

After completing all conditions, the research team asked the participants to leave feedback about their overall experience or other comments they thought might be useful. At that point, the research team provided answers to the participants' questions, such as details of the study, and asked the participants about their user experiences. None of our participants dropped out and needed less than 1 hour to complete the study.

4 Result

For our statistical analyses, we used the self-reported ratings, the completion time, and the number of puzzle pieces of the logged data as dependent variables and the five experimental conditions as independent variables (see Section 3.3 for the experimental conditions). We analyzed the previously mentioned data with one-way repeated measures ANOVA with *post hoc* Bonferroni correction for multiple comparisons. We analyzed the gaze data using a two-way repeated measures ANOVA with *post hoc* Bonferroni correction following a 5 (Conditions: LI vs. LSC vs. MSC vs. HSC vs. HI) × 3 (Gazes: virtual character [VC] vs. puzzle pieces [PP] vs. puzzle goal [PG]) factorial design. The normality assumptions were validated with Q–Q plots of the residuals.

4.1 Self-Reported Ratings

We provide descriptive statistics of our self-reported ratings and the patterns of difference across the examined conditions for each measurement in Table 1.

Perceived Intelligence. We found a significant effect of the self-correction behavior (Wilks' $\Lambda = .161$, $F[4, 19] = 24.807$, $\eta_p^2 = .839$, $p = .000$) across the five conditions. The *post hoc* pairwise comparison indicated that in the LI condition ($M = 2.16$, $SD = .30$), our participants rated the virtual character's perceived intelligence lower than in the MSC condition ($M = 3.70$, $SD = .30$; $p = .000$), HSC condition ($M = 4.78$, $SD = .24$; $p = .000$), and HI condition ($M = 5.76$, $SD = .17$; $p = .000$). Moreover, participants in the LSC condition ($M = 2.64$, $SD = .29$) rated the virtual character's perceived intelligence significantly lower than in the MSC condition ($p = .011$), HSC condition ($p = .000$), and HI condition ($p = .000$). Participants in the MSC condition rated the virtual character's perceived intelligence significantly lower than in the HSC condition ($p = .011$) and HI condition ($p = .000$). Additionally, participants in the HSC condition rated the virtual character's perceived intelligence lower than in the HI condition ($p = .005$). However, our participants did not report a significant difference between the LI and LSC conditions ($p = .150$).

Intelligence Comparison. The statistical analysis revealed a significant effect of the self-correction behavior (Wilks' $\Lambda = .290$, $F[4, 19] = 11.614$, $\eta_p^2 = .710$, $p = .000$) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition ($M = 1.35$, $SD = .18$) rated their intelligence comparison lower than in the HSC condition ($M = 2.78$, $SD = .35$; $p = .021$) and HI condition ($M = 3.83$, $SD = .38$; $p = .000$). Moreover, participants in the LSC condition ($M = 1.17$, $SD = .10$) rated their intelligence comparison significantly lower than in the MSC condition ($M = 1.96$, $SD = .23$; $p = .022$), HSC condition ($p = .002$), and HI condition ($p = .000$). Participants in the MSC condition rated their intelligence comparison lower than in the HI condition ($p = .000$). Additionally, participants in the HSC condition rated their intelligence comparison lower than in the HI condition ($p = .017$). However, we did not find a significant difference between the LI and LSC conditions ($p = 1.000$) and between the MSC and HSC conditions ($p = .166$).

Table 1. Descriptive Statistics of Perceived Intelligence, Intelligence Comparison, Virtual Character's Awareness (Private, Public, and Surroundings), Trust, Performance, Enjoyment, Frustration, and Desire for Future Interaction

Perceived Intelligence

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	2.16	.30	1.00	7.00	(LI, LSC) < MSC < HSC < HI
LSC	2.64	.29	1.00	6.00	
MSC	3.70	.30	1.33	6.17	
HSC	4.78	.24	2.00	7.00	
HI	5.76	.17	4.17	7.00	

Intelligence Comparison

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	1.35	.18	1.00	5.00	LI < (HSC, HI)
LSC	1.17	.10	1.00	3.00	LSC < (MSC, HSC, HI)
MSC	1.96	.23	1.00	5.00	(MSC,HSC) < HI
HSC	2.78	.35	1.00	6.00	
HI	3.83	.38	1.00	7.00	

Private Awareness

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	2.02	.30	1.00	6.00	(LI, LSC) < (MSC, HSC) < HI
LSC	2.24	.25	1.00	6.00	
MSC	3.37	.32	1.00	6.00	
HSC	3.70	.31	1.00	7.00	
HI	4.72	.32	1.50	7.00	

Public Awareness

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	1.76	.26	1.00	5.50	LI < (MSC, HSC) < HI
LSC	2.17	.26	1.00	5.00	LSC < (HSC, HI)
MSC	3.02	.32	1.00	6.00	
HSC	3.33	.32	1.00	6.50	
HI	4.20	.35	1.00	7.00	

Surrondings Awareness

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	2.13	.32	1.00	7.00	(LI, LSC) < (MSC, HSC, HI)
LSC	2.35	.32	1.00	7.00	
MSC	3.46	.36	1.00	6.00	
HSC	3.74	.41	1.00	7.00	
HI	4.33	.37	1.00	7.00	

Trust

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	2.17	.23	1.00	5.50	LI < (MSC, HSC) < HI
LSC	2.48	.22	1.00	5.00	LSC < HSC < HI
MSC	3.21	.21	1.00	4.75	
HSC	3.49	.24	1.00	5.50	
HI	4.33	.28	1.50	7.00	

Performance

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	1.52	.20	1.00	5.00	(LI, LSC) < MSC < HSC < HI
LSC	1.83	.22	1.00	5.00	
MSC	3.22	.30	1.00	6.00	
HSC	4.57	.26	2.00	7.00	
HI	6.00	.19	4.00	7.00	

Enjoyment

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	2.39	.41	1.00	7.00	(LI, LSC) < (MSC, HSC, HI)
LSC	2.48	.36	1.00	7.00	MSC < HI
MSC	4.09	.42	1.00	7.00	
HSC	5.04	.35	1.00	7.00	
HI	5.30	.28	3.00	7.00	

Frustration

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	5.22	.42	1.00	7.00	(HSC, HI) < LI
LSC	5.39	.35	1.00	7.00	(MSC, HSC, HI) < LSC
MSC	3.70	.42	1.00	7.00	HI < MSC
HSC	3.09	.38	1.00	7.00	
HI	1.87	.28	1.00	6.00	

Desire for Future Interaction

Condition	*M*	*SD*	*Min*	*Max*	Pattern of Difference
LI	2.48	.41	1.00	7.00	(LI, LSC) < (MSC, HSC) < HI
LSC	2.35	.36	1.00	7.00	
MSC	4.00	.40	1.00	7.00	
HSC	4.65	.37	1.00	7.00	
HI	5.70	.24	3.00	7.00	

We report the M, SD, minimum (Min), maximum (Max), and patterns of differences. LI, 0% Intelligence without Self-correction; LSC, 0% Intelligence with 0% Accurate Self-correction; MSC, 0% Intelligence with 50% Accurate Self-correction; HSC, 0% Intelligence with 100% Accurate Self-correction; HI, 100% Intelligence without Self-correction.

Virtual Character's Private Awareness. There was a significant effect of the self-correction behavior (Wilks' $\Lambda = .260$, $F[4, 19] = 13.492$, $\eta_p^2 = .740$, $p = .000$) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition ($M = 2.02$, $SD = .30$) rated their virtual character's private awareness lower than in the MSC condition ($M = 3.37$, $SD = .32$;

p = .017), HSC condition (M = 3.70, SD = .31; p = .001), and HI condition (M = 4.72, SD = .32; p = .000). Participants in the LSC condition (M = 2.24, SD = .25) rated their virtual character's private awareness lower than in the MSC condition (p = .010), HSC condition (p = .001), and HI condition (p = .000). Moreover, participants in the MSC condition rated their virtual character's private awareness lower than in the HI condition (p = .019). Additionally, participants in the HSC condition rated their virtual character's private awareness lower than in the HI condition (p = .029). However, we did not find a significant difference between the LI and LSC conditions (p = 1.000) and between the MSC and HSC conditions (p = 1.000).

Virtual Character's Public Awareness. We found a significant effect of the self-correction behavior (Wilks' Λ = .281, $F[4, 19]$ = 12.166, η_p^2 = .719, p = .000) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition (M = 1.76, SD = .26) rated their virtual character's public awareness lower than in the MSC condition (M = 3.02, SD = .32; p = .003), HSC condition (M = 3.33, SD = .32; p = .000), and HI condition (M = 4.20, SD = .35; p = .000). In the LSC condition (M = 2.17, SD = .26), our participants rated their virtual character's public awareness lower than in the HSC condition (p = .005) and HI condition (p = .000). Moreover, participants in the MSC condition rated their virtual character's public awareness lower than in the HI condition (p = .008). Additionally, participants in the HSC condition rated their virtual character's public awareness lower than in the HI condition (p = .021). However, we did not find a significant difference between the LI and LSC conditions (p = .184), between the LSC and MSC conditions (p = .185) and between the MSC and HSC conditions (p = 1.000).

Virtual Character's Surroundings Awareness. Our statistical analysis revealed a significant effect of the self-correction behavior (Wilks' Λ = .369, $F[4, 19]$ = 8.107, η_p^2 = .631, p = .001) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition (M = 2.13, SD = .32) rated their virtual character's surroundings awareness lower than in the MSC condition (M = 3.46, SD = .36; p = .005), HSC condition (M = 3.74, SD = .41; p = .001), and HI condition (M = 4.33, SD = .37; p = .000). Moreover, participants in the LSC condition (M = 2.35, SD = .32) rated their virtual character's surroundings awareness lower than in the MSC condition (p = .037), HSC condition (p = .016), and HI condition (p = .000). However, we did not find a significant difference between the LI and LSC conditions (p = 1.000), between the MSC and HSC conditions (p = 1.000), between the MSC and HI conditions (p = .124), and between the HSC and HI conditions (p = .394).

Trust. There was a significant effect of the self-correction behavior (Wilks' Λ = .344, $F[4, 19]$ = 9.044, η_p^2 = .656, p = .000) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition (M = 2.17, SD = .23) rated their trust lower than in the MSC condition (M = 3.21, SD = .21; p = .002), HSC condition (M = 3.49, SD = .24; p = .003), and HI condition (M = 4.33, SD = .28; p = .000). Additionally, participants in the LSC condition (M = 2.48, SD = .22) rated their trust lower than in the HSC condition (p = .013) and HI condition (p = .000). Moreover, participants in the MSC condition rated their trust significantly lower than in the HI condition (p = .001), and participants in the HSC condition rated their trust lower than in the HI condition (p = .009). Finally, we did not find a significant difference between the LI and LSC conditions (p = .884), between the LSC and MSC conditions (p = .053), and between the MSC and HSC conditions (p = 1.000).

Performance. The statistical analysis revealed a significant effect of the self-correction behavior (Wilks' Λ = .059, $F[4, 19]$ = 75.095, η_p^2 = .941, p = .000) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition (M = 1.52, SD = .20) rated the virtual character's performance lower than in the MSC condition (M = 3.22, SD = .30; p = .000), the HSC condition (M = 4.57, SD = .26; p = .000), and the HI condition (M = 6.00, SD = .19; p = .000). Additionally, participants in the LSC condition (M = 1.83, SD = .22) rated the virtual character's performance lower than in the MSC condition (p = .001), HSC condition (p = .000), and HI condition (p = .000). Moreover, participants in the MSC condition rated the virtual character's performance lower than in the HSC condition (p = .000) and the HI condition (p = .000). Finally, participants in the HSC condition rated the virtual character's performance lower than in the HI condition (p = .000). However, we did not find a significant difference between the LI and LSC conditions (p = .897).

Enjoyment. There was a significant effect of the self-correction behavior (Wilks' Λ = .239, $F[4, 19] = 15.135$, $\eta_p^2 = .761$, p = .000) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition (M = 2.39, SD = .41) rated their enjoyment lower than in the MSC condition (M = 4.09, SD = .42; p = .004), HSC condition (M = 5.04, SD = .35; p = .000), and HI condition (M = 5.30, SD = .28; p = .000). Moreover, participants in the LSC condition (M = 2.48, SD = .36) rated their enjoyment lower than in the MSC condition (p = .011), HSC condition (p = .000), and HI condition (p = .000). Additionally, participants in the MSC condition rated their enjoyment lower than in the HI condition (p = .039). However, we did not find a significant difference between the LI and LSC conditions (p = 1.000), between the MSC and HSC conditions (p = .055), and between the HSC and HI conditions (p = 1.000).

Frustration. We found a significant effect of the self-correction behavior (Wilks' Λ = .281, $F[4, 19] = 12.140$, $\eta_p^2 = .719$, p = .000) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition (M = 5.22, SD = .42) rated their frustration higher than in the HSC condition (M = 3.09, SD = .38; p = .011) and HI condition (M = 1.87, SD = .28; p = .000). Participants in the LSC condition (M = 5.39, SD = .35) rated their frustration higher than in the MSC condition (M = 3.70, SD = .42; p = .020), HSC condition (p = .002), and HI condition (p = .000). Moreover, participants in the MSC condition rated their frustration higher than in the HI condition (p = .007). However, we did not find significant differences between the LI and LSC conditions (p = 1.000), between the LI and MSC conditions (p = .096), between the MSC and HSC conditions (p = 1.000), and between the HSC and HI conditions (p = .079).

Desire for Future Interaction. There was a significant effect of the self-correction behavior (Wilks' Λ = .241, $F[4, 19] = 14.992$, $\eta_p^2 = .759$, p = .000) across the five conditions. The *post hoc* pairwise comparison indicated that participants rated the LI condition (M = 2.48, SD = .41) lower than the MSC condition (M = 4.00, SD = .40; p = .010), HSC condition (M = 4.65, SD = .37; p = .000), and HI condition (M = 5.70, SD = .24; p = .000). Participants in the LSC condition (M = 2.35, SD = .36) rated their desire for future interaction lower than in the MSC condition (p = .003), HSC condition (p = .000), and HI condition (p = .000). Moreover, participants in the MSC condition rated their desire for future interaction lower than in the HI condition (p = .009). Additionally, participants in the HSC condition rated their desire for future interaction lower than in the HI condition (p = .034). However, we did not find a significant difference between the LI and LSC conditions (p = 1.000) and between the MSC and HSC conditions (p = .610).

4.2 Logged Data

We provide descriptive statistics and the patterns of difference across the examined conditions of the dwell gazing data in Table 2, and completion time and number of pieces in Table 3.

Table 2. Detailed Results of Our Study for the Gazing Data

	Virtual Character				**Puzzle Pieces**				**Puzzle Goal**				
	M	SD	Min	Max	M	SD	Min	Max	M	SD	Min	Max	Pattern of Difference
LI	.14	.10	.01	.40	.10	.12	.00	.49	.02	.08	.00	.25	
LSC	.14	.11	.00	.36	.09	.11	.01	.39	.04	.10	.00	.12	
MSC	.15	.10	.00	.37	.07	.08	.00	.30	.03	.08	.02	.33	
HSC	.18	.09	.01	.45	.08	.07	.00	.28	.01	.03	.00	.27	
HI	.15	.08	.00	.35	.07	.08	.02	.36	.03	.08	.00	.40	
Main Effect (Conditions)													
F	.450												
p	.771												
η_p^2	.087												
Main Effect (Gazes)													
F	**11.840**												(PP, PG) < VC
p	**.000**												
η_p^2	**.530**												
Interaction Effect (Conditions×Gazes)													
F	1.285												
p	.321												
η_p^2	.407												
Conditions df = 4 (Error df = 19), Gazes df = 2 (Error df = 21), and Interaction df = 8 (Error df = 15).													

We report the M, SD, minimum (Min), maximum (Max), and patterns of differences. We present significant results with bold font. HI, 100% Intelligence without Self-correction; HSC, 0% Intelligence with 100% Accurate Selfcorrection; LI, 0% Intelligence without Self-correction; LSC, 0% Intelligence with 0% Accurate Self-correction; MSC, 0% Intelligence with 50% Accurate Self-correction; P G, puzzle goal; P P, puzzle pieces; VC, virtual character.

Table 3. Descriptive Statistics of Completion Time and Number of Puzzle Pieces Placed by Participants

Completion Time						**Number of Puzzle Pieces**					
Condition	M	SD	Min	Max	Pattern of Difference	Condition	M	SD	Min	Max	Pattern of Difference
LI	193.50	15.62	64.597	347.278	(LI, LSC, MSC) < (HSC, HI)	LI	25.00	.00	25.00	25.00	HI < HSC < MSC < (LI = LSC)
LSC	227.37	33.30	71.361	827.542		LSC	25.00	.00	25.00	25.00	
MSC	148.63	10.95	78.750	268.917		MSC	19.78	2.29	16.00	23.00	
HSC	105.08	5.81	69.889	174.556		HSC	16.21	3.14	9.00	21.00	
HI	93.41	4.64	54.556	156.944		HI	13.26	3.22	5.00	19.00	

We report the M, SD, minimum (Min), maximum (Max), and patterns of differences. HI, 100% Intelligence without Self-correction; HSC, 0% Intelligence with 100% Accurate Self-correction; LI, 0% Intelligence without Self-correction; LSC, 0% Intelligence with 0% Accurate Self-correction; MSC, 0% Intelligence with 50% Accurate Self-correction.

Dwell Gazing. We did not find a statistically significant result on dwell gazing data for the Conditions factor (Wilk's Λ = .913, $F[4, 19]$ = .450, p = .771, η_p^2 = .087). However, the simple main effect analysis on the Gazes factor indicated a statistically significant result (Wilk's Λ = .470, $F[2, 21]$ = 11.840, p = .000, η_p^2 = .530). The *post hoc* pairwise comparison indicated that participants gazed at the virtual character more time (M = .15, SE = .02) than the puzzle goal (M = .03, SE = .01; p = .000) and puzzle pieces (M = .08, SE = .02; p = .009). However, we did not find a statistically significant result for the Conditions × Gazes interaction (Wilk's Λ = .593, $F[8, 15]$ = 1.285, p = .321, η_p^2 = .407).

Completion Time. The statistical analysis revealed a significant effect of the self-correction behavior (Wilks' $\Lambda = .229$, $F[4,19] = 16.032$, $\eta_p^2 = .771$, $p = .000$) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition ($M = 193.50$, $SD = 15.62$) spent more time than in the HSC condition ($M = 105.08$, $SD = 5.81$; $p = .000$) and HI condition ($M = 93.41$, $SD = 4.64$; $p = .000$). Additionally, participants in the LSC condition ($M = 227.37$, $SD = 33.30$) spent more time than when we exposed them to the HSC condition ($p = .011$) and HI condition ($p = .002$). Moreover, participants in the MSC condition ($M = 148.63$, $SD = 10.95$) spent more time than in the HSC condition ($p = .001$) and HI condition ($p = .000$). However, we did not find significant differences between the LI and LSC conditions ($p = 1.000$), between the LI and MSC conditions ($p = .100$), between the LSC and MSC conditions ($p = .127$), and between the HSC and HI conditions ($p = .690$).

Number of Puzzle Pieces. The statistical analysis showed a significant effect (Wilks' $\Lambda = .066$, $F[4,19] = 94.163$, $\eta_p^2 = .934$, $p = .000$) across the five conditions. The *post hoc* pairwise comparison indicated that participants in the LI condition ($M = 25.00$, $SD = .00$) placed more puzzle pieces than in the MSC condition ($M = 19.78$, $SD = 2.29$; $p = .000$), HSC condition ($M = 16.21$, $SD = 3.14$; $p = .000$), and HI condition ($M = 13.26$, $SD = 3.22$; $p = .000$). Additionally, participants in the LSC condition ($M = 25.00$, $SD = .00$) placed more puzzle pieces than in the MSC condition ($p = .000$), HSC condition ($p = .000$), and HI condition ($p = .000$). Moreover, participants in the MSC condition placed more puzzle pieces than in the HSC condition ($p = .00$) and HI condition ($p = .000$). Finally, participants in the HSC condition placed more puzzle pieces than in the HI condition ($p = .000$).

4.3 Qualitative Data

After our study participants had completed all conditions, we collected their impressions of our VR application and interactions with the virtual character. We grouped the data into categories concerning the intelligence of the virtual character and their interactions with the virtual character and our VR application.

Participants commented that they noticed the different behaviors of the virtual character in different conditions. Specifically, P7 wrote: "The first couple tries [LI and HSC] was kinda interesting to see how the other individual would interact with the puzzle, the last one [HI] felt like it was just cheating and knew where each piece would go in the beginning." P8 stated: "There was one experiment where it was actually doing very well, putting the edge pieces in first, and was pretty accurate. But immediately after that, it looked like it was doing it randomly again..." P11 mentioned: "From games 2 [HSC] and 3 [MSC], there was a shift in her being faster at interacting with the puzzle and placing it in the correct spots." Finally, P18 reported: "Some of the individuals were intelligent enough to collaborate, but on the other hand, some of them were not intelligent..."

Moreover, some participants reported their observations of the virtual character. Specifically, P4 reported: "The sounds of the virtual person make me think that she's confident at what she's doing, but for some scenarios, she keeps placing the wrong puzzle." Moreover, P11 mentioned: "The last game [LSC], I noticed she was able to turn the puzzle piece, which surprised me."

According to the collected comments, it can be said that most participants enjoyed the VR puzzle co-solving experience with the virtual character, and some participants reported their preferred condition. Specifically, P2 wrote: "Great study!," Both P13 and P5 stated: "It was good," and P7 wrote: "It was a fun experience in general." Moreover, P8 reported: "This was interesting, though. I did like seeing how she was processing where to put the pieces." P9 wrote: "This was a nice experience," and P16 reported: "It was a great experience, I enjoyed playing in the last condition [HSC]." For P18, "It was interesting that we can play jigsaw puzzles with a virtual character... Overall, I enjoyed solving a jigsaw puzzle with the virtual character." P19 said: "I prefer the first one [HI] and last one [HSC]...," while P20 stated: "It was very enjoyable, I am really fond of VR, and seeing this only brings me hope for the future." P22 similarly mentioned that "It was really fun playing jigsaw with individuals that were intelligent...," and P23 wrote: "This was my first time using a VR device, and it was fantastic!"

However, some participants stated that it was frustrating when they solved the puzzle with an unintelligent virtual character, such as in the conditions of 0% Intelligence without Self-correction or 0% Intelligence with 0% Self-correction. Specifically, P1 mentioned: "It's frustrating that sometimes the virtual character does not do their part." For P12, "The experience of the last condition [LI] is bad since the other individual keeps doing the wrong thing and didn't notice she did the thing wrong." P18 reported: "Some of them were not intelligent, so I had to do it again, which was kind of burdensome." P19 mentioned that "The rest of them [LI, LSC, and MSC] are really boring and tough," while P22 said it was "...very frustrating when it came to those who would just put down pieces randomly."

5 Discussion

We asked our participants to provide self-reported ratings on co-solving a jigsaw puzzle with a virtual character to understand how the self-correction behaviors assigned to the virtual character affected their perceptions and experiences. We also collected gaze and completion time data to understand how participants observed the virtual character, the tasks they had to work on, and how fast they solved the puzzle. The statistical analysis revealed several interesting findings, which we discuss in the following subsections.

5.1 RQ1: Intelligence

Perceived intelligence concerns how humans perceive the intelligence of a system [90]. Several researchers have investigated the factors of perceived intelligence, such as anthropomorphism [49], animacy [6], and understandability [24]. We conducted this study because we wanted to extend current knowledge and explore how study participants perceived the intelligence of a virtual character when we assigned different levels of intelligence and self-correction behavior to it.

We found significant differences in the results of perceived intelligence. When either the *Intelligence* or the *Accuracy of Self-correction* increased, the perceived intelligence ratings of our participants also increased. Based on the patterns of differences (see Table 1), we found significant results between the LSC, MSC, and HSC conditions, indicating that the *Accuracy of Self-correction* improved the perceived intelligence partially when the virtual character made mistakes. We think that this finding indicates that the *Accuracy of Self-correction* can be another factor impacting the perceived intelligence of virtual characters along with animacy [6], anthropomorphism [49], appearance [17], and understandability [24]. However, we should also consider task complexity, which researchers defined by various components, such as the number of elements [96], relationships between tasks [97], time pressure [34], and cognitive demands [4]. For example, if the task is complex and challenging for the participant, the self-correction and errors made by the virtual character may not be perceived in the same way. In such scenarios, the ability of the virtual character to self-correct could be seen as a more critical and valuable trait, as the perceived task complexity is highly related to the cognitive workload [86]; thus, potentially enhancing its perceived intelligence even more significantly.

We also found significant results—similar to those for perceived intelligence—when examining the intelligence comparison ratings. We found that the *Intelligence* parameter was more affected than the *Accuracy of Self-correction* from the comparison between the HSC and HI conditions. We also found that, except for the HI condition, participants provided ratings below the scale's mean

(< 3.5). This finding agrees with Ullman et al. [91] and Bennet et al. [7], as both studies reported lower ratings for robots when their participants compared intelligence between themselves and robots. It should be noted that even if the rating in the HI condition was above the scale's average (participants provided an $M = 3.83$), we cannot really argue that they rated the virtual character as more intelligent than themselves. This means that participants, even if they interacted with a highly intelligent character that was able to solve the jigsaw puzzle efficiently, indicated that the virtual character's intelligence was not enough to make them rate it as a truly highly intelligent creature.

One possible explanation is that the virtual character could be perceived as "smarter" than the participants in the specific context of solving the puzzle but not "smarter" than them in general. Our participants might have recognized the virtual character's proficiency in the narrow task of jigsaw puzzle co-solving without extending that recognition to a broader, more general intelligence. Additionally, our participants may have inherently considered their own intelligence as more comprehensive, involving emotional understanding, creativity, and adaptability across different interaction scenarios, which a virtual character's task-specific competence does not encapsulate. This distinction highlights the multifaceted nature of intelligence. It suggests that task-specific capabilities do not necessarily translate to a perception of overall higher intelligence [31, 85], as human intelligence is not only about problem-solving skills but also includes creative, practical, and emotional aspects, which are often absent in virtual characters [23].

5.2 RQ2: Awareness

Following the definition of awareness given in Govern's and Marsch's [33] study, we defined a virtual character's awareness as how much they understand the virtual environment. In our study, we focused on the virtual character's awareness of its inner feelings (private awareness), awareness of the participant (public awareness), and awareness of how the puzzle-solving progressed (surroundings awareness). We observed that in all the examined awareness ratings, the means of the LI, LSC, and MSC conditions were below the scale's mean (< 3.5). Such low ratings indicated that the virtual character did not convince our participants it was aware of them, aware of the jigsaw puzzle co-solving process, or even aware of itself since it could not solve the puzzle independently.

Our intention behind examining the virtual character's self-correction behavior was to make the virtual character behave less like a robot and more like a human, as humans tend to make mistakes, recognize them, and subsequently adjust their actions accordingly [71]. Becoming aware of the task requirements and dynamically correcting its actions would allow the virtual character to perform more intuitively and fluidly, closely mimicking human behavior. From the virtual character's private awareness result, we found that the rating of the HI condition was significantly higher than the other conditions. We can argue that such a result was because our participants thought the virtual character should not make mistakes if it were aware of itself. We interpret this finding according to Nirenburg et al. [70], who mentioned that imitating human behavior could improve self-awareness. Thus, we think the HI condition had a higher rating than the HSC condition because the virtual character solved the puzzle efficiently and in an error-free way, like the participants. There was also a significant result between the MSC and HSC groups of conditions and the LI and LSC groups of conditions. We found that the participants provided higher ratings on the MSC and HSC groups of conditions. This result extends the findings of a previous study of the self-awareness of a humanoid robot [73], indicating that focusing on the inner state and self-modifying the representation makes a robot self-aware. We think self-correction improved the participants' perception of the virtual character's private awareness.

In the result concerning the virtual character's public awareness, we found that the rating of the HI condition was significantly higher than the other conditions. Our finding agrees with Hayes et al. [37], who indicated that their study participants thought the robot understood the dance movements they taught it well until it made the first mistake. Our study participants felt the virtual character was not conscious of our participants when it made its first mistake, although it could fix it. We also found the rating of the LSC condition was significantly lower than that of the HSC conditions. Again, this finding extends Hayes et al.'s study, which reported that repeated mistakes by a robot could invoke negative feedback. We thus argue that our study participants felt the virtual character was not conscious of our participants when it tried to correct its mistake but still repeatedly placed the puzzle piece in the wrong spot.

Finally, the virtual character's surroundings awareness showed different results than the other types of awareness. We found the LI and LSC groups of conditions had lower ratings than the MSC, HSC, and HI groups of conditions. This finding extends Drury et al.'s [25] study, which defined a human's awareness of the overall goals of a task with robots as one of the components of the perception of the environment in human–robot interaction. We think that our study participants felt the virtual character was aware of the goal of the jigsaw puzzle co-solving process when it could solve the puzzle like themselves, and our virtual character invoked higher ratings of the surroundings awareness during the MSC, HSC, and HI conditions.

5.3 RQ3: User Experience

To explore how *Intelligence* and *Accuracy of Self-correction* affect user experiences, we included items in our questionnaire to measure trust, performance, enjoyment, frustration, and desire for future interaction. We discuss our findings in the following paragraphs.

The result for trust showed that our participants rated the HI condition highest while also indicating that there were differences from the other conditions to which we exposed them. Based on this finding, we can argue that the mistakes of the virtual character negatively affected study participants' trust ratings, even if the virtual character was able to correct its mistakes. We base this interpretation of our results on previous studies that explored trust in human–robot interaction. Specifically, Hald et al. [36] reported that although their robot could fix its mistakes, the trust of their study participants was already broken. Similarly, Roesler et al. [79] indicated that their study participants' trust in the robot decreased after it made errors. Furthermore, Salem et al. [80] found that study participants reported higher ratings on trust when the robot followed user input correctly than when it behaved incorrectly. However, we also found a significant difference between the LSC and HSC conditions. Our finding extends Hald et al.'s [36] study. Such a significant result indicates that self-correction accuracy could partially recover trust. This finding aligns with another previous study on the correlation between verbal communication mistakes and trustworthiness [87], in which the authors stated that the mistakes of virtual humans decrease their trustworthiness but temporarily. Additionally, we should note that we observed a p-value at the border of significance between the LSC and MSC conditions ($p = .053$). While this borderline statistical significance may be a consequence of multiple comparisons, it does raise the possibility that trust could be further enhanced under conditions of increased levels of *Accuracy of Self-correction*. Overall, we can argue that because the reduced trustworthiness is temporary, the virtual character can only recover the trust partially by fixing its mistake.

Regarding performance rating, our participants rated the HI condition higher. At the same time, our participants were also able to identify and report differences with other conditions. Our finding agrees with Hald et al.'s [36] study, which reported that people rated the performance of a robot that made a mistake lower than one that did not. This finding also extends the study of Esterwood et al. [26], which reported that any verbal repair strategies from the robots for their multiple mistakes did not positively impact study participants' perceptions of their performances. We think our study participants provided higher scores on the HI condition than on other conditions because error-free decisions are more important than self-corrections.

The enjoyment results showed significant differences between the two groups of conditions; one group comprises the LI and LSC conditions, and the other comprises the MSC, HSC, and HI conditions. We found that when the virtual character was able to solve the puzzle independently, such behavior improved the enjoyment of our participants. We interpret this finding based on a previous study concerning enjoyment in games [9], which reported that a user's enjoyment decreased when the user felt a high responsibility to the virtual character. Moreover, we observed a p-value at the border of significance between the MSC and HSC conditions ($p = .055$). Such a borderline result should be interpreted with caution, as it does not denote a statistical significance. However, we think it suggests that even small improvements in the levels of *Accuracy of Self-correction* could potentially influence participants' enjoyment levels. All in all, we argue the participants felt high responsibility and less enjoyment when fixing the virtual character's mistakes.

The frustration results showed that participants rated the LSC condition higher than the LI condition. This finding confirms Cho's [15] study, which indicated people felt frustrated when a virtual assistant misunderstood a question and gave a wrong answer. The virtual character in the LSC condition made mistakes repeatedly and caused more frustration than the virtual character that did not try to self-correct its mistakes.

Finally, we asked our participants to report their desire for future interactions and whether they are willing to interact in the future with the different behaviors assigned to the virtual character. Our participants rated the HI condition the highest. We also found significant results between the HI and the other conditions. Our finding agrees with and extends the study of Cuadra et al. [19], which reported that participants preferred a perfect voice assistant to a voice assistant that carried out irrelevant tasks according to users' commands and then corrected them. We think the participants preferred an error-free virtual character to a virtual character that makes mistakes, even if the virtual character can correct itself.

5.4 RQ4: Behavioral Responses

Unfortunately, we could not find significant results in any of the collected dwell gazing measurements across the five experimental conditions. Thus, we cannot argue that the *Intelligence* or the *Accuracy of Self-correction* of the virtual character impacted the gaze of our participants. Perhaps this could be due to the pseudo-gazing methodology not providing precise tracked-gaze data. However, as we discuss later, visual attention needs to be reexamined to provide clearer conclusions. Nevertheless, the gaze factor revealed that participants gazed at the virtual character more than the puzzle goal and puzzle pieces. This finding suggests that the virtual character was the leading actor in the interaction scenario we developed. Therefore, our participants' visual attention was primarily drawn to the virtual character's actions and behaviors rather than the specific task conditions. Thus, we attribute this finding to the engaging nature of the virtual character, which likely drew participants' attention regardless of the examined conditions.

We found significant results in the completion time measurement between the group composed of the LI, LSC, and MSC conditions and the group composed of the HSC and HI conditions. We also found that the more intelligent the virtual character became, or the more accurate its self-correction was, the fewer puzzle pieces our participants needed to place to complete the puzzle. Based on these findings, we can argue that a more intelligent virtual character could indeed help participants co-solve a problem faster and significantly reduce their workload.

5.5 Limitations

When implementing virtual characters, it is necessary to consider their functionalities and user experiences. Although all participants experienced our VR jigsaw puzzle co-solving experience without any issues, we would like to report several limitations. It should be noted that these limitations do not invalidate our implementation and study; instead, they provide guides for and improve subsequent research.

First, the participants selected answers from a UI menu using the VR controller instead of verbal communication. Although we did not collect data related to participants' level of immersion, we think that such an interaction mechanism might have impacted our participants' immersion. Thus, we argue that further development of and experimentation with speech-based communication will help us improve our participants' levels of immersion.

Second, our virtual character exhibited fixed loop-based behavior with limited variations, impacting the realism of its actions. Although self-correction behaviors were integrated, certain conditions required the virtual character to deliberately place puzzle pieces incorrectly on the first attempt to demonstrate self-correction. This repetitive error reduced the virtual character's realism. Furthermore, the virtual character's finger animations were not active, leading to less natural and believable interactions. To enhance realism, future implementations should incorporate a more sophisticated control over the frequency and nature of self-correction behaviors and activate detailed finger animations to create more lifelike interactions.

Third, we consider the inability of our virtual character to convince our study participants that it is aware of the virtual environment, the task, and itself as an additional limitation. Unfortunately, we did not implement events such as a phone ring or a fly that buzzes and the corresponding reactions/animations that could make the virtual character behave as if it were more aware of the environment in which the co-solving process is situated. Moreover, our dialogs were short and not highly related to the performance of our virtual character. Thus, such short dialogs made our participants think the virtual character was unaware of the task. We thus argue that such additions could enhance study participants' perception of a virtual character's awareness.

Fourth, we think including haptic feedback might have improved the overall user experience and interaction realism [46]. However, we did not implement haptic feedback in our study. While it could potentially enhance the experience by mimicking the tactile feedback humans rely on when solving a real jigsaw puzzle, the current limitations of VR controllers in providing realistic haptic sensations made its effectiveness uncertain [21]. Additionally, to our knowledge, there is no prior research on representing correct and incorrect puzzle piece placement using haptic feedback. Thus, we argue that additional research in this direction is needed.

Fifth, we acknowledge that our study participants were young, which may have influenced the results. Consequently, our findings might not apply to older adults or other age groups. We consider it an important area for future research to investigate whether similar results would be observed in experiments involving older adults.

Finally, we used a point-of-view method to check what the user gazed at while solving the puzzle. Although this method provided some data about where our study participants were gazing (i.e., what was in the center of their fields of view), such data did not reflect our participants' actual gaze or fixations. We think an eye-tracker would enable us to collect more reliable data that we can use to understand how participants co-solved the puzzle and interacted with the virtual character.

6 Conclusions and Future Work

Several researchers have explored how the behavior assigned to virtual characters can impact human perception. Although several studies have investigated interaction with virtual characters, we have limited knowledge of the impact of the self-correction behavior of a virtual character on humans' perceptions and user experiences. Therefore, in this article, we explored how the self-correction behavior of a virtual character impacted our study participants. We implemented a virtual character that could co-solve a jigsaw puzzle and self-correct its mistakes. Then, we asked participants to report how they perceived the virtual character and their user experience, and we collected application logs to understand how they interacted with the virtual character. The statistical analysis showed that the self-correction behavior impacted our participants' perception of the virtual character and their user experiences. Also, we found our participants gazed at the virtual character more than the puzzle pieces and puzzle goal in all experimental conditions.

Although this study provides noteworthy results, it also has limitations, such as the absence of verbal communication with the virtual character. We think addressing the mentioned limitations will help us expand the findings we report in this article. Therefore, in future work, we would like to integrate text-to-speech and speech-to-text into the virtual character to enable verbal communication and a chatbot based on language models to generate a dialog to enhance the virtual character's communication abilities and potential awareness. Moreover, we would like to explore how different behaviors (e.g., selfish or competitive behavior) could impact how participants perceive and interact with the virtual character.

Appendix

A Survey

We developed a survey to understand how the self-correction behavior of a virtual character affects human perception and user experiences. The survey comprises 21 items examining ten variables: perceived intelligence, intelligence comparison, virtual character's awareness (private awareness, public awareness, and surroundings awareness), trust, performance, enjoyment, frustration, and desire for future interaction. We provide our survey along with the anchors of the scales in Table A1.

Table A1. The Survey We Used in Our Study

#	Item	Anchors of the Scale
	Perceived Intelligence (Moussawi and Koufaris [66])	
Q1	The other individual was able to operate without my intervention.	1 = Never, 7 = Always
Q2	The other individual was aware of the virtual environment.	1 = Never, 7 = Always
Q3	The other individual was able to set and pursue tasks by herself in anticipation of future needs.	1 = Never, 7 = Always
Q4	The other individual was able to complete tasks quickly.	1 = Never, 7 = Always
Q5	The other individual was able to find and process the necessary information for completing the task.	1 = Never, 7 = Always
Q6	The other individual was able to adapt/adjust its behavior based on prior events.	1 = Never, 7 = Always
	Intelligence Comparison	
Q7	Do you think the other individual was smarter than you?	1 = Not at all, 7 = Totally
	Virtual Character's Private Awareness (Govern and Marsch [33])	
Q8	The other individual was conscious of her actions.	1 = Not at all, 7 = Totally
Q9	The other individual was aware of her innermost actions.	1 = Not at all, 7 = Totally
	Virtual Character's Public Awareness (Govern and Marsch [33])	
Q10	The other individual was concerned about the way we played the jigsaw puzzle.	1 = Not at all, 7 = Totally
Q11	The other individual was self-conscious about the way we played the jigsaw puzzle.	1 = Not at all, 7 = Totally

(Continued)

Table A1. Continued

#	Item	Anchors of the Scale
	Virtual Character's Surroundings Awareness (Govern and Marsch [33])	
Q12	The other individual was aware of everything in the virtual environment.	1 = Not at all, 7 = Totally
Q13	The other individual was conscious of what was going on around it.	1 = Not at all, 7 = Totally
	Trust (Jian et al. [43])	
Q14	I am suspicious of the other individual's intention.	1 = Not at all, 7 = Totally
Q15	I am confident in the other individual.	1 = Not at all, 7 = Totally
Q16	The other individual is dependable.	1 = Not at all, 7 = Totally
Q17	The other individual is reliable.	1 = Not at all, 7 = Totally
	Performance	
Q18	Rate the performance of the other individual.	1 = Not good, 7 = Very good
	Enjoyment	
Q19	Did you enjoy solving the jigsaw puzzle with the other individual?	1 = Not at all, 7 = Totally
	Frustration	
Q20	I felt frustrated when interacting with the other individual.	1 = Not at all, 7 = Totally
	Desire for Future Interaction	
Q21	Are you willing to interact with the other the other individual again?	1 = Not at all, 7 = Totally

References

[1] Pedro Acevedo, Alejandra Magana, Christos Mousas, Yoselyn Walsh, Hector Will Pinto, and Bedrich Benes. 2022. Effects of tactile feedback on conceptual understanding of electromagnetism in a virtual reality experience. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct '22)*. IEEE, 588–593.

[2] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2571–2582.

[3] Andreas Aristidou and Joan Lasenby. 2011. Fabrik: A fast, iterative solver for the inverse kinematics problem. *Graphical Models* 73, 5 (2011), 243–260.

[4] Nathan R. Bailey and Mark W. Scerbo. 2007. Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 321–348.

[5] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. arXiv:1909.07528. Retrieved from https://doi.org/10.48550/arXiv.1909.07528

[6] Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. 2009. Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics* 1 (2009), 195–204.

[7] Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '17)*. IEEE, 6589–6594.

[8] Timothy Bickmore and Justine Cassell. 1999. Small talk and conversational storytelling in embodied conversational interface agents. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*, 87–92.

[9] Nicholas David Bowman, Mary Beth Oliver, Ryan Rogers, Brett Sherrick, Julia Woolley, and Mun-Young Chung. 2016. In control or in their shoes? How character attachment differentially influences video game enjoyment and appreciation. *Journal of Gaming & Virtual Worlds* 8, 1 (2016), 83–99.

[10] Jeffrey M. Bradshaw, Paul Feltovich, and Matthew Johnson. 2017. Human-agent interaction. In *The Handbook of Human-Machine Interaction*. CRC Press, 283–300.

[11] Catherina Burghart, Christian Gaertner, and Heinz Woern. 2006. Cooperative solving of a children's jigsaw puzzle between human and robot: First results. In *Proceedings of the Cognitive Robotics: Papers from the AAAI Workshop: Papers from the 2006 AAAI Workshop*. M. Beetz, K. Rajan, M. Thielscher, and R. B. Rusu (Eds.). Citeseer, 33–39.

[12] Aaron Burns, Ben Sugden, Laura Massey, and Tom Salter. 2019. Gaze-based object placement within a virtual reality environment, September 17 2019. US Patent 10,416,760.

[13] Marc Cavazza, Fred Charles, and Steven J. Mead. 2001. Agents' interaction in virtual storytelling. In *Proceedings of the Intelligent Virtual Agents: Third International Workshop (IVA '01)*. Springer, 156–170.

[14] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2014. How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits. In *Proceedings of the Human Behavior Understanding: 5th International Workshop (HBU '14)*. Springer, 1–15.

[15] Janghee Cho. 2018. Mental models and home virtual assistants (HVAs). In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6.

[16] Minsoo Choi, Yeling Jiang, Farid Breidi, Christos Mousas, and Mesut Akdere. 2022. A mixed reality platform for collaborative technical assembly training. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, 1–2.

[17] Minsoo Choi, Alexandros Koilias, Matias Volonte, Dominic Kao, and Christos Mousas. 2023. Exploring the appearance and voice mismatch of virtual characters. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 555–560.

[18] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

[19] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. In *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1), 1–24.

[20] Dixuan Cui, Dominic Kao, and Christos Mousas. 2021. Toward understanding embodied human-virtual character interaction through virtual and tactile hugging. *Computer Animation and Virtual Worlds* 32 (3–4), e2009.

[21] Dixuan Cui and Christos Mousas. 2022. Estimating the just noticeable difference of tactile feedback in oculus quest 2 controllers. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '22)*. IEEE, 1–7.

[22] Sylvain Daronnat. Human-agent trust relationships in a real-time collaborative game. In *Proceedings of the Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, 18–20.

[23] Kerstin Dautenhahn. 1998. The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied Artificial Intelligence* 12, 7–8 (1998), 573–617.

[24] Amol Deshmukh, Bart Craenen, Mary Ellen Foster, and Alessandro Vinciarelli. 2018. The more I understand it, the less I like it: The relationship between understandability and godspeed scores for robotic gestures. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '18)*. IEEE, 216–221.

[25] Jill L. Drury, Jean Scholtz, and Holly A. Yanco. 2003. Awareness in human-robot interactions. In *Proceedings of the SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance*, Vol. 1. IEEE, 912–918.

[26] Connor Esterwood and Lionel P. Robert. 2021. Do you still trust me? Human-robot trust repair strategies. In *Proceedings of the 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN '21)*. IEEE, 183–188.

[27] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (2009), 1149–1160.

[28] Sara Salevati Feldman, Ozge Nilay Yalcin, and Steve DiPaola. 2017. Engagement with artificial intelligence through natural interaction models. In *Proceedings of the Electronic Visualisation and the Arts (EVA '17)*. BCS Learning & Development, 296–303.

[29] Patrick Fissler, Olivia Caroline Küster, Daria Laptinskaya, Laura Sophia Loy, Christine A. F. Von Arnim, and Iris-Tatjana Kolassa. 2018. Jigsaw puzzling taps multiple cognitive abilities and is a potential protective factor for cognitive aging. *Frontiers in Aging Neuroscience* 10 (2018), 299.

[30] Piotr Fratczak, Yee Mey Goh, Peter Kinnell, Laura Justham, and Andrea Soltoggio. 2021. Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics* 82 (2021), 103078.

[31] Howard E. Gardner. 2011. *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.

[32] Manuel Giuliani and Alois Knoll. 2011. Evaluating supportive and instructive robot roles in human-robot interaction. In *Proceedings of the Social Robotics: 3rd International Conference (ICSR '11)*. Springer, 193–203.

[33] John M. Govern and Lisa A. Marsch. 2001. Development and validation of the situational self-awareness scale. *Consciousness and Cognition* 10, 3 (2001), 366–378.

[34] Frank L. Greitzer. 2005. Toward the development of cognitive task difficulty metrics to support intelligence analysis research. In *Proceedings of the 4th IEEE Conference on Cognitive Informatics (ICCI '05)*. IEEE, 315–320.

[35] Siqi Guo, Minsoo Choi, Dominic Kao, and Christos Mousas. 2024. Collaborating with my doppelgänger: The effects of self-similar appearance and voice of a virtual character during a jigsaw puzzle co-solving task. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 7, 1 (2024), 1–23.

[36] Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. "An error occurred!"-trust repair with virtual robot using levels of mistake explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, 218–226.

[37] Cory J. Hayes, Maryam Moosaei, and Laurel D. Riek. 2016. Exploring implicit human responses to robot mistakes in a learning from demonstration task. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '16)*. IEEE, 246–252.

[38] Thomas T. Hewett, Ronald Baecker, Stuart Card, Tom Carey, Jean Gasen, Marilyn Mantei, Gary Perlman, Gary Strong, and William Verplank. 1992. *ACM SIGCHI Curricula for Human-Computer Interaction*. ACM, 1–162.

[39] Jon-Chao Hong, Ming-Yueh Hwang, Ker-Ping Tam, Yi-Hsuan Lai, and Li-Chun Liu. Effects of cognitive style on digital jigsaw puzzle performance: A gridware analysis. *Computers in Human Behavior* 28, 3 (2012), 920–928.

[40] Wei-Chia Huang, Sai-Keung Wong, Matias Volonte, and Sabarish V. Babu. 2023. Impact of socio-demographic attributes and mutual gaze of virtual humans on users' visual attention and collision avoidance in VR. *IEEE Transactions on Visualization and Computer Graphics* 30, 9 (2023), 6146–6163.

[41] Kiran Ijaz, Anton Bogdanovych, and Simeon Simo. 2011. Enhancing the believability of embodied conversational agents through environment-, self- and interaction-awareness. In *Proceedings of the Conferences in Research and Practice in Information Technology Series*, 107–116.

[42] Xing-Da Jhan, Sai-Keung Wong, Elham Ebrahimi, Yuwen Lai, Wei-Chia Huang, and Sabarish V. Babu. 2022. Effects of small talk with a crowd of virtual humans on users' emotional and behavioral responses. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3767–3777.

[43] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.

[44] Dominic Kao, Alejandra J. Magana, and Christos Mousas. 2021. Evaluating tutorial-based instructions for controllers in virtual reality games. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–28.

[45] Alexandros Koilias, Christos Mousas, and Christos-Nikolaos Anagnostopoulos. 2019. The effects of motion artifacts on self-avatar agency. In *Informatics* 6, MDPI (2019), 18.

[46] Alexandros Koilias, Christos Mousas, and Christos-Nikolaos Anagnostopoulos. 2020. I feel a moving crowd surrounds me: Exploring tactile feedback during immersive walking in a virtual crowd. *Computer Animation and Virtual Worlds* 31, 4–5 (2020), e1963.

[47] Claudia Krogmeier, Christos Mousas, and David Whittinghill. 2019. Human–virtual character interaction: Toward understanding the influence of haptic feedback. *Computer Animation and Virtual Worlds* 30, 3–4 (2019), e1883.

[48] Walter Lasecki and Jeffrey Bigham. 2012. Self-correcting crowds. In *Proceedings of the CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 2555–2560.

[49] Jae-Gil Lee, Ki Joon Kim, Sangwon Lee, and Dong-Hee Shin. 2015. Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction* 31, 10 (2015), 682–691.

[50] Michael Lewis. 1998. Designing for human-agent interaction. *AI Magazine* 19, 2 (1998), 67–67.

[51] Huimin Liu, Minsoo Choi, Dominic Kao, and Christos Mousas. 2023. Synthesizing game levels for collaborative gameplay in a shared virtual environment. *ACM Transactions on Interactive Intelligent Systems* 13, 1 (2023), 1–36.

[52] Huimin Liu, Minsoo Choi, Liuchuan Yu, Alexandros Koilias, Lap-Fai Yu, and Christos Mousas. 2022. Synthesizing shared space virtual reality fire evacuation training drills. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct '22)*. IEEE, 459–464.

[53] Huimin Liu, Zhiquan Wang, Angshuman Mazumdar, and Christos Mousas. 2021. Virtual reality game level layout design for real environment constraints. *Graphics and Visual Computing* 4 (2021), 200020.

[54] Kuan-Yu Liu, Matias Volonte, Yu-Chun Hsu, Sabarish V Babu, and Sai-Keung Wong. 2019. Interaction with proactive and reactive agents in box manipulation tasks in virtual environments. *Computer Animation and Virtual Worlds* 30, 3–4 (2019), e1881.

[55] Yarden Livnat, James Agutter, Shaun Moon, and Stefano Foresti. 2005. Visual correlation for situational awareness. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '05)*. IEEE, 95–102.

[56] Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2018. Culture, errors, and rapport-building dialogue in social agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 51–58.

[57] Amama Mahmood, Jeanie W. Fung, Isabel Won, and Chien-Ming Huang. 2022. Owning mistakes sincerely: Strategies for mitigating AI errors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–11.

[58] David G. McNeely-White, Francisco R. Ortega, J. Ross Beveridge, Bruce A. Draper, Rahul Bangar, Dhruva Patil, James Pustejovsky, Nikhil Krishnaswamy, Kyeongmin Rim, Jaime Ruiz, Isaac Wang. 2019. User-aware shared perception for embodied agents. In *Proceedings of the IEEE International Conference on Humanized Computing and Communication (HCC)*. IEEE, 46–51.

[59] Tim Merritt, Kevin McGee, Teong Leong Chuah, and Christopher Ong. 2011. Choosing human team-mates: Perceived identity as a moderator of player preference and enjoyment. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, 196–203.

[60] Tim Merritt, Christopher Ong, Teong Leong Chuah, and Kevin McGee. 2011. Did you notice? Artificial team-mates take risks for players. In *Proceedings of the Intelligent Virtual Agents: 10th International Conference (IVA '11)*. Springer, 338–349.

[61] Chenlin Ming, Jiacheng Lin, Pangkit Fong, Han Wang, Xiaoming Duan, and Jianping He. 2023. Hicrisp: A hierarchical closed-loop robotic intelligent self-correction planner. arXiv:2309.12089. Retrieved from https://doi.org/10.48550/arXiv.2309.12089

[62] Hazel Morton and Mervyn A. Jack. 2005. Scenario-based spoken interaction with virtual agents. *Computer Assisted Language Learning* 18, 3 (2005), 171–191.

[63] Christos Mousas. 2018. Performance-driven dance motion control of a virtual partner character. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 57–64.

[64] Christos Mousas, Dimitris Anastasiou, and Ourania Spantidi. 2018. The effects of appearance and motion of virtual characters on emotional reactivity. *Computers in Human Behavior* 86 (2018), 99–108.

[65] Christos Mousas, Alexandros Koilias, Dimitris Anastasiou, Banafsheh Rekabdar, and Christos-Nikolaos Anagnostopoulos. 2019. Effects of self-avatar and gaze on avoidance movement behavior. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR '19)*. IEEE, 726–734.

[66] Sara Moussawi and Marios Koufaris. 2019. Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale development and validation. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 115–124.

[67] Arslan Munir, Alexander Aved, and Erik Blasch. Situational awareness: Techniques, challenges, and prospects. *AI* 3, 1 (2022), 55–77.

[68] Luis Muñoz-Saavedra, Lourdes Miró-Amarante, and Manuel Domínguez-Morales. Augmented and virtual reality evolution and future tendency. *Applied Sciences* 10, 1 (2020), 322.

[69] Michael G. Nelson, Alexandros Koilias, Dominic Kao, and Christos Mousas. 2023. Effects of speed of a collocated virtual walker and proximity toward a static virtual character on avoidance movement behavior. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '23)*. IEEE, 930–939.

[70] Sergei Nirenburg, Marjorie McShane, and Stephen Beale. 2010. Aspects of metacognitive self-awareness in Maryland virtual patient. In *Proceedings of the AAAI Fall Symposium Series,* 69–74.

[71] Donald A. Norman. 1981. Categorization of action slips. *Psychological Review* 88, 1 (1981), 1.

[72] Donald A. Norman. 1994. How might people interact with agents. *Communications of the ACM* 37, 7 (1994), 68–71.

[73] Rony Novianto and Mary-Anne Williams. 2009. The role of attention in robot self-awareness. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '09)*. IEEE, 1047–1053.

[74] David Obremski, Jean-Luc Lugrin, Philipp Schaper, and Birgit Lugrin. 2021. Non-native speaker perception of intelligent virtual agents in two languages: The impact of amount and type of grammatical mistakes. *Journal on Multimodal User Interfaces* 15 (2021), 229–238.

[75] Sai Krishna Pathi, Annica Kristoffersson, Andrey Kiselev, and Amy Loutfi. 2019. F-formations for social interaction in simulation using virtual agents and mobile robotic telepresence systems. *Multimodal Technologies and Interaction* 3, 4 (2019), 69.

[76] Muhammad Hasham Qazi and Muhammad Palize Qazi. Introducing vaifu: A virtual agent for introducing and familiarizing users in VR. In *Proceedings of the 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET '23)*. IEEE, 1–6.

[77] Chao Qu, Willem-Paul Brinkman, Yun Ling, Pascal Wiggers, and Ingrid Heynderickx. 2014. Conversations with a virtual human: Synthetic emotions and human responses. *Computers in Human Behavior* 4 (2014), 58–68.

[78] Jeff Rickel and W. Lewis Johnson. 2000. Task-oriented collaboration with embodied agents in virtual worlds. In *Embodied Conversational Agents*. J. Cassell, J. Sullivan, and S. Prevost (Eds.), MIT Press, 95–122.

[79] Eileen Roesler, Linda Onnasch, and Julia I. Majer. 2020. The effect of anthropomorphism and failure comprehensibility on human-robot trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64, SAGE Publications Sage CA, Los Angeles, CA, 107–111.

[80] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 141–148.

[81] Satragni Sarkar, Dejanira Araiza-Illan, and Kerstin Eder. 2017. Effects of faults, experience, and personality on trust in a robot co-worker. arXiv:1703.02335. Retrieved https://doi.org/10.48550/arXiv.1703.02335

[82] Glenda L. Satne. 2014. Interaction and self-correction. *Frontiers in Psychology* 5 (2014), 798.

[83] Richard Skarbez, Aaron Kotranza, Frederick P. Brooks, Benjamin Lok, and Mary C. Whitton. 2011. An initial exploration of conversational errors as a novel method for evaluating virtual human experiences. In *Proceedings of the IEEE Virtual Reality Conference*. IEEE, 243–244.

[84] Travis Steel, Dane Kuiper, and R. Z. Wenkstern. 2010.Context-aware virtual agents in open environments. In *Proceedings of the 2010 6th International Conference on Autonomic and Autonomous Systems*. IEEE, 90–96.

[85] Robert J. Sternberg. 1985. *Beyond IQ: A Triarchic Theory of Human Intelligence*. CUP Archive, 1985.

[86] Gerald Stollnberger, Astrid Weiss, and Manfred Tscheligi. 2013. "The harder it gets" exploring the interdependency of input modalities and task complexity in human-robot collaboration. In *Proceedings of the IEEE RO-MAN*. IEEE, 264–269.

[87] Jacob Stuart, Karen Aul, Michael D. Bumbach, Anita Stephen, Alexandre Gomes De Siqueira, and Benjamin Lok. 2022. The effect of virtual humans making verbal communication mistakes on learners' perspectives of their credibility, reliability, and trustworthiness. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR '22)*. IEEE, 455–463.

[88] William R. Swartout, Jonathan Gratch, Randall W. Hill Jr, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. Toward virtual humans. *AI Magazine* 27, 2 (2006), 96–96.

[89] Ning Tan, Gaëtan Pruvost, Matthieu Courgeon, Céline Clavel, Yacine Bellik, and Jean-Claude Martin. 2011. A location-aware virtual character in a smart room: Effects on performance, presence and adaptivity. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 399–02.

[90] Serge Thill, Maria Riveiro, and Maria Nilsson. 2015. Perceived intelligence as a factor in (semi-)autonomous vehicle UX. In *Proceedings of the "Experiencing Autonomous Vehicles: Crossing the Boundaries between a Drive and a Ride" Workshop in Conjunction (CHI '15)*.

[91] Daniel Ullman, Lolanda Leite, Jonathan Phillips, Julia Kim-Cohen, and Brian Scassellati. 2014. Smart human, smarter robot: How cheating affects perceptions of social agency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 36, 2996–3001.

[92] Matias Volonte, Yu-Chun Hsu, Kuan-Yu Liu, Joe P. Mazer, Sai-Keung Wong, and Sabarish V. Babu. 2020. Effects of interacting with a crowd of emotional virtual humans on users' affective and non-verbal behaviors. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR '20)*. IEEE, 293–302.

[93] Matias Volonte, Eyal Ofek, Ken Jakubzak, Shawn Bruner, and Mar Gonzalez-Franco. 2022. Headbox: A facial blendshape animation toolkit for the microsoft rocketbox library. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW '22)*. IEEE, 39–42.

[94] Yuqiong Wang, Peter Khooshabeh, and Jonathan Gratch. 2013. Looking real and making mistakes. In *Proceedings of the Intelligent Virtual Agents: 13th International Conference (IVA '13)*. Springer, 339–348.

[95] Evan James Williams. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry* 2, 2 (1949), 149–168.

[96] Terry M. Williams. 1999. The need for new paradigms for complex projects. *International Journal of Project Management* 17, 5 (1999), 269–273.

[97] Robert E. Wood. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37, 1 (1986), 60–82.

[98] Biao Xie, Huimin Liu, Rawan Alghofaili, Yongqi Zhang, Yeling Jiang, Flavio Destri Lobo, Changyang Li, Wanwan Li, Haikun Huang, Mesut Akdere, Christos Mousas, and Lap-Fai Yu. A review on virtual reality skill training applications. *Frontiers in Virtual Reality* 2 (2021), 645153.

[99] Holly A. Yanco and Jill Drury. 2004. Where am I?" acquiring situation awareness using a remote robot platform. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3. IEEE, 2835–2840.

[100] Zi-Ming Ye, Jun-Long Chen, Miao Wang, and Yong-Liang Yang. 2021. Paval: Position-aware virtual agent locomotion for assisted virtual reality navigation. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '21)*. IEEE, 239–247.

[101] Zhenjie Zhao and Xiaojuan Ma. 2020. Situated learning of soft skills with an interactive agent in virtual reality via multimodal feedback. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 25–27.

Received 5 December 2023; revised 19 May 2024; accepted 22 July 2024