HSVRS: A Virtual Reality System of the Hide-and-Seek Game to Enhance Gaze Fixation Ability for Autistic Children

Chengyan Yu, Shihuan Wang, Dong Zhang, Yingying Zhang, Chaoqun Cen, Zhixiang You, Xiaobing Zou, Hongzhu Deng, Ming Li *Senior Member, IEEE*

***Abstract*—Numerous children diagnosed with Autism Spectrum Disorder (ASD) exhibit abnormal eye gaze pattern in communication and social interaction. In this study, we aim to investigate the effectiveness of the Hide and Seek Virtual Reality System (HSVRS) in improving gaze fixation abilities in children with Autism Spectrum Disorder (ASD). Our hypothesis is that engaging in a hide-and-seek game within a virtual environment, particularly with a customized avatar resembling a familiar figure, would significantly enhance gaze fixation skills compared to traditional interventions without supplementary VR intervention. Thirty-six children with ASD were involve in this pilot study in three groups: the avatar customized group, the avatar uncustomized group and the control group. The control group only received human intervention while the avatar groups received additional VR-assisted interventions. The effect of HSVRS was measured by a six-point Likert subjective questionnaire and demonstrated significant improvements in gaze fixation abilities in the VR-assisted intervention groups compared to the control group (P=0.006, 0.001). Moreover, the Avatar customized group, which interacted with a familiar-looking avatar, obtained noticeable increments in gaze fixation metrics (P=0.036, 0.005, 0.001). Experimental results show the effectiveness of utilizing VR technology to complement regular interventions in terms of improving gaze fixation abilities for young children with ASD.**

***Index Terms*—Autism, virtual reality, digital intervention, voice conversion, gaze tracking**

I. Introduction

AUTISM Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by impairment in communication and social interaction, restricted interests and stereotyped behaviors [1], [2]. The prevalence of ASD has substantially increased in recent decades [3]. Currently, the overall prevalence of ASD among eight-year-old children in the United States has been reported to be approximately 1 in 44 children (2.27%) [4]. Abnormal eye gaze patterns are early emerging symptoms of ASD [5], and they have been incorporated into many diagnostic criteria and instruments. Previous studies have shown that individuals with ASD tend to exhibit reduced gaze attention towards the upper face region, which contains social and emotional information [6], and increased gaze attention towards non-social objects or items. Previous eye-tracking studies have also shown that children with ASD are more prone to gaze fixation deficits compared to typically developing individuals and those with other developmental disabilities, thus highlighting abnormal gaze patterns as critical characteristics of ASD [7], [8].

The existing solutions, e.g., Applied Behavior Analysis (ABA) [9] and the Early Start Denver Model (ESDM) [10], improve the gaze fixation skills/abilities of ASD children with behavioral interventions. These methods focus on building social skills through structured play and interaction with therapists and caregivers. ABA is a systematic and evidence-based approach that applies principles of behaviorism to improve socially significant behaviors. ABA focuses on understanding and modifying behaviors through the analysis of antecedents, behaviors, and consequences. The process involves breaking down complex skills into smaller, manageable tasks, and using positive reinforcement to encourage desired behaviors while employing strategies to decrease challenging behaviors. The ESDM is an early intervention approach designed for children with or at risk for ASD. It integrates a developmental and relationship-based framework, incorporating behavioral and developmental principles. ESDM is delivered in a play-based, naturalistic setting and emphasizes joint attention, social engagement, and communication skills. While these interventions have shown positive outcomes, they typically involve manual training and require substantial human resources and financial costs [11]. Due to the limited number of professional therapists specializing in ASD, children with ASD sometimes miss crucial opportunities for improvement due to long waiting times for intervention, particularly in developing countries where resources are scarce [12]. In this context,

Chengyan Yu and Shihuan Wang contributed equally to this work.

This research is funded in part by the Science and Technology Program of Guangzhou City (202007030011), National Natural Science Foundation of China (62171207, 81873801, 62173353) and DKU Synear and Wang-Cai Seed Grant. (Corresponding author: Hongzhu Deng, Ming Li)

This work involved human subjects in its research. This study was approved by the Third Affiliated Hospital of Sun Yat-sen University Institutional Review Board (IRB No. [2020]02-118-03) and Duke Kunshan University Institutional Review Board (IRB No. 2022ML065).

Chengyan Yu and Dong Zhang are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: yuchy33@mail2.sysu.edu.cn; zhangd@mail.sysu.edu.cn).

Ming Li is with the Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Data Science Research Center, Duke Kunshan University, Kunshan, China, 215316, (e-mail: ming.li369@dukekunshan.edu.cn).

Shihuan Wang, Yingying Zhang, Chaoqun Cen, Xiaobing Zou, Hongzhu Deng are with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, No.600 Tianhe Road, Guangzhou, China, 510630, (email: wangshh53@mail.sysu.edu.cn; summer09251228@163.com; cenchq@mail.sysu.edu.cn; zouxb@vip.tom.com; denghzh@mail.sysu.edu.cn).

Zhixiang You is with JazzVision Tech Co, Science and Technology City, Nanhu District, Jiaxing, Zhejiang, China, 314016, (email: zhx.you@gmail.com)

Fig. 1. The virtual environment in the HSVRS game is set in a room in which the parent avatar can hide in various locations. The parent avatar in this screenshot is hiding in front of the bed, waving at the user.

digital applications hold significant potential to enhance social interaction and facilitate daily tasks for individuals with ASD in a cost-effective manner [13]

VR technology provides a unique set of affordances that can be highly beneficial for learners, especially those with special needs such as ASD. The immersive nature of VR allows children to engage in a 360-degree environment that simulates real-world scenarios, offering a safe and controlled space for learning and practice. This immersive experience helps children with ASD better understand social cues and engage in social interactions, which are often challenging for them. The multisensory aspect of VR enables the stimulation of multiple senses simultaneously, e.g. sight, sound and even movement. For children with ASD, this facilitates a more comprehensive learning experience. While VR shows promise, other alternatives such as augmented reality (AR) and computer-based games have also been explored for similar purposes. However, VR's immersive nature and the ability to track and adapt to the user's gaze provide unique advantages for auxiliary ASD invervention. Virtual reality (VR) technology with eye tracker has been increasingly applied in intervention for children with ASD [14]. The eye tracker in the VR headset can serve as a control device, replacing the need for a regular joystick and thereby reducing interaction challenges. While virtual reality-based intervention cannot replace interventions conducted by professional therapists, it can provide a secure and immersive virtual environment in which children with ASD can engage with systems to enhance their social skills [15]. The structured space within VR is another critical affordance that aids learning for children with ASD. The controlled environment allows for a systematic introduction of learning objectives and a gradual increase in complexity, accommodating the child's learning pace and needs. Moreover, we can also design certain high-risk skill learning sessions in the virtual environment, e.g. crossing the road. Furthermore, VR interventions can be customized to address sensory sensitivities frequently observed in individuals with ASD, creating a more comfortable learning environment. The ability to hear and respond to a familiar voice, as achieved through our deepfake technology, can further engage the child and promote positive interactions within the VR environment [16].

In recent years, there has been a growing interest among researchers in utilizing virtual reality devices for assistive interventions in autism. Some researchers have focused on enhancing gaze fixation abilities [17], [18], while others aim to improve joint attention skills [19]–[22], social communication skills [23]–[25], emotional development [26], [27], among other areas of investigation. However, the issue of transferability and the effectiveness of digital interventions in enhancing real-life skills among autistic children has indeed been a significant topic of discussion in prior literature, often highlighted as a vital area for future research [28], [29]. Recently, an evidence-based systematic review [30] shows that there are moderate evidence about the effectiveness of VR-based treatments in ASD. For example, [25], [31] used professional behavior assessment tools, e.g. PEP-3, ABAS-II, in the pre- and post-assessments which reflect the child's practical application skills and adaptive behavioral skills. Moreover, the safety and feasibility of VR intervention were analyzed and discussed in [32], results show that immersive VR is safe, feasible, and usable for adolescents with ASD.

The aforementioned studies demonstrate that advanced virtual reality technology has the potential to assist in the treatment of autism. However, only a limited number of studies have integrated VR with cutting-edge technologies, including deep learning-based computer vision and voice manipulation, to enhance conventional interventions. Previous research has indicated the importance of personalization in VR interventions for typical individuals [33], [34]. It has emphasized the positive impacts of customizable virtual environments and avatars, particularly in therapeutic contexts. Similar conclusions have been drawn in the field of autism treatment [35], [36]. The previous studies often faced challenges related to the inconvenience and costliness of customization. Therefore, our study focuses on leveraging deep learning-based computer vision and voice manipulation techniques to enable affordable customization of game characters within virtual reality devices. To the best of our knowledge, our work is the first to adopt deep-learning based deepfake technology to clone real person in the VR-assisted ASD intervention.

Deepfake technology refers to using artificial intelligence, specifically deep learning, to create or manipulate multimedia content to portray something that does not exist in reality. In the context of our research, we employ deep-learning-based deepfake techniques to customize the appearance and voice of virtual avatars within the Hide and Seek Virtual Reality System (HSVRS). This customization uses a participant's photographs and audio recordings to generate a virtual representation that closely resembles a familiar figure, e.g., a parent, of the child with Autism Spectrum Disorder (ASD). Our utilization of deepfake technology is based on the premise that personalization can significantly enhance the engagement and therapeutic outcomes of virtual reality interventions for children with ASD. By creating a more immersive and real virtual environment, we aim to better improve the child's gaze fixation abilities, which are often impaired in ASD and can impact social interactions.

In this study, we present a novel VR-based system called the Hide and Seek Virtual Reality System (HSVRS). The

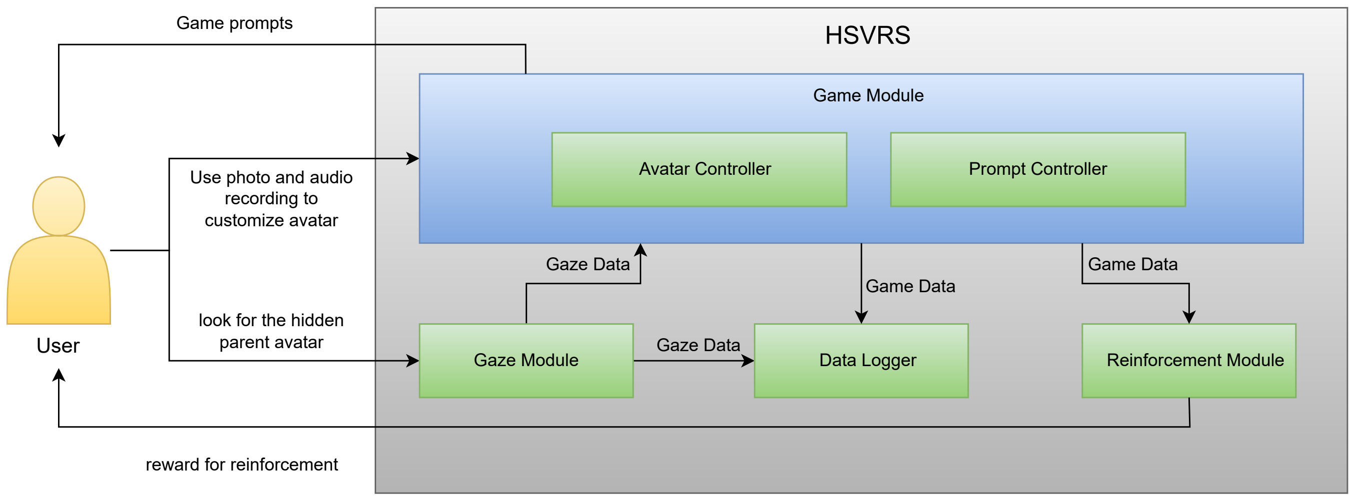


Fig. 2. Overview of the HSVRS system. Our HSVRS system consists of four components: Game Module (including Avatar Controller and Prompt Controller), Gaze Module, Data Logger and Reinforcement Module.

system is specifically designed to facilitate hide-and-seek games for children with Autism Spectrum Disorder (ASD) by utilizing customized virtual avatars, such as their parents. The primary aim of HSVRS is to improve the eye gaze ability of children with ASD through engaging gameplay. HSVRS provides a virtual home scene equipped with furniture, creating a suitable environment for playing hide-and-seek. During gameplay, users simply need to turn their heads and utilize their eye gaze to locate the hidden avatar in order for the game to function properly. The system incorporates built-in eye-tracking devices, making it accessible and user-friendly for children of all age groups. Moreover, the avatar within the system can be customized by uploading a facial photograph and a few audio recordings of the parent. This customization feature allows the avatar to closely resemble and sound like the child's actual parent. Consequently, this study focuses on two pivotal research questions. 1) How effective is the proposed HSVRS in enhancing gaze fixation abilities in children with ASD? 2) Does the ability to customize avatars within the HSVRS impact engagement levels and intervention outcomes in comparison to the use of non-customized avatars?

II. System Design

We design the Hide and Seek Virtual Reality System (HSVRS) to help children practice their skills of gaze fixation. The central gameplay of HSVRS revolves around a hide-and-seek game, wherein the user interacts with an avatar representing their parents. Previous research has demonstrated the engaging nature of hide-and-seek games for children with Autism Spectrum Disorder (ASD) [37]. Our system features a virtual family room scene with furniture, providing a realistic environment for the parent avatars to conceal themselves. Fig. 1 presents a screenshot captured during actual gameplay, offering a glimpse into the immersive experience. Playing hide-and-seek with VR headsets is straightforward as users can actively participate by controlling their eye gaze. A notable aspect of HSVRS is its capacity to customize the parent avatars using cutting-edge deep learning-based computer vision and voice conversion techniques. By providing a single photo and a few audio recordings, we can ensure that the avatar closely resembles the child's parent both in appearance and voice.

A. Overview

Fig. 2 presents the interaction diagrams depicting the communication between the user and the components of HSVRS. The system comprises four key components: the Game Module (consisting of the Avatar Controller and Prompt Controller), Gaze Module, Data Logger, and Reinforcement Module. Within the Game Module, the Avatar Controller utilizes a parent's photo and 80 audio utterances to customize the avatar. It manages the parent avatars, facilitating their hiding and transitioning between different hiding locations. Concurrently, the Prompt Controller receives gaze data from the Gaze Module, enabling it to provide prompts when the user struggles to locate the parent avatar. The collaborative efforts of the Avatar Controller and Prompt Controller ensure the smooth progression and functionality of the game. The Gaze Module, equipped with a built-in eye tracker, captures the user's gaze data during the hide-and-seek gameplay. It relays this data to both the Game Module and the Data Logger for further processing and analysis. The Data Logger receives and stores both gaze data and game data, enabling subsequent analysis and evaluation. The Reinforcement Module leverages the game data to determine whether the user has successfully found the parent avatar. In the event of a successful find, the Reinforcement Module provides rewards in the form of visual and auditory reinforcers. Visual reinforcers encompass captivating special effects, thumbs up gestures from the parents' avatars, and interactions with user-favorite cartoon characters. Auditory reinforcers include pleasant music and encouraging words spoken by the parent character, among other positive stimuli.

B. Game Module

The Game Module comprises two main components: the Avatar Controller and the Prompt Controller. This module

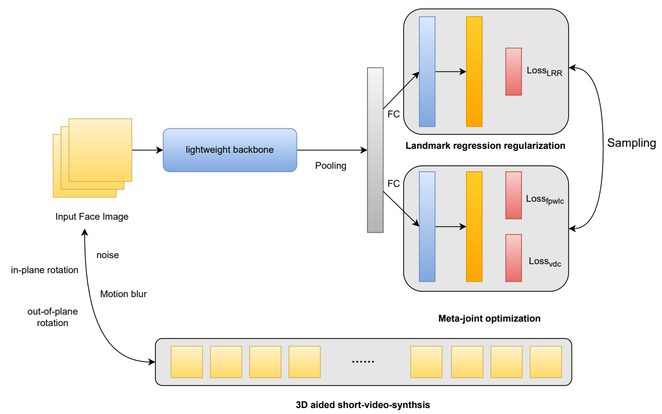


Fig. 3. The architecture of the 3DDFA_V2 network.The 3DDFA_V2 architecture consists of four parts: the lightweight backbone, the landmark-regression regularization, the meta-joint optimization of fWPDC and VDC [38], [39], and the 3D-aided short-video-synthesis.

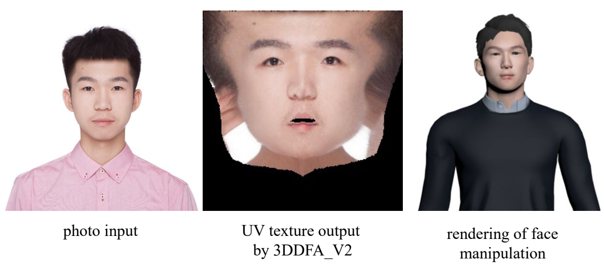


Fig. 4. An example of face manipulation using the first author's face.

assumes responsibility for managing the hiding behavior of the parent avatar within the virtual home scene. It facilitates the hiding process and dynamically alters the avatar's hiding location once it has been discovered. Additionally, in cases where the user struggles to locate the parent avatar, the Prompt Controller intervenes by offering helpful prompts to assist the user in the gameplay.

1) Avatar Controller: In order to better understanding the deepfake techonology, we introduce two definitions.

- Deep Learning: Deep Learning is a subset of machine learning that teaches computers to learn from experience. It uses neural networks with many layers (hence 'deep') to analyze various factors and make decisions or predictions.
- Deepfake: Deepfake is a synthetic media that portrays the appearance of a person using deep learning techniques. This technology has been used in our study to create a personalized avatar that looks and sounds like a child's parent, making the virtual interaction more familiar and engaging.

The key elements of the deepfake techonology in our research are the input (photographs and audio recordings), the process (deep learning algorithms), and the output (customized avatars).

We employed the application 3D Studio Max for designing the animation of the parent avatars. In order to examine the potential impact of face and voice deepfake manipulation on

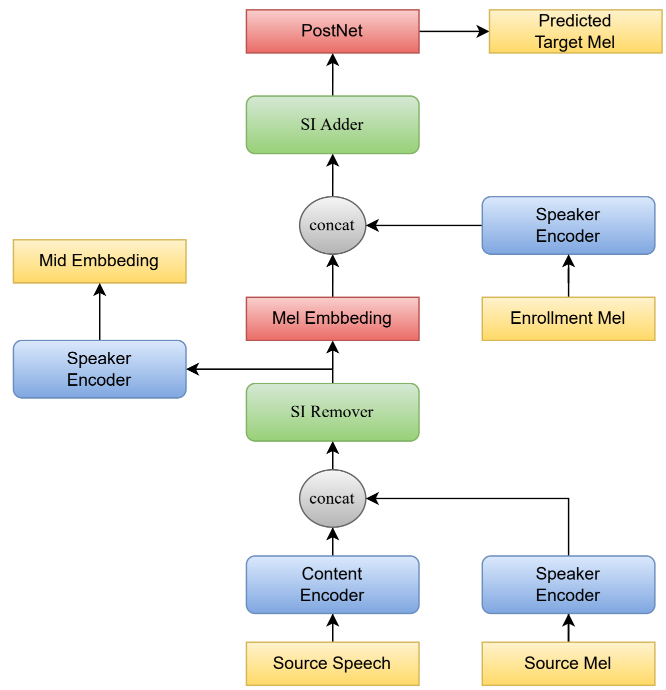


Fig. 5. The structure of the SIG-VC system, where SI stands for speaker information [40]. SIG-VC proposes a supervised intermediate representation to reduce speaker information and obtain purer content representations.

the intervention effect, we created two distinct types of avatars: simulated avatars and customized avatars. The simulated avatars feature a consistent face and voice across all instances. On the other hand, for the customized avatars, we utilized the photos and recordings provided by the participants to manipulate both the appearance and voices of the avatars. This allowed us to create personalized avatars that closely resemble and sound like the participants' actual parents.

We now describe the implementation details in the face manipulation module. We pre-designed some of the rough facial shapes, such as the large eyes, high nose, and tiny mouth. Moreover, we crafted a range of common hairstyles such as the rounded inch, the backward style, and the bald head. In 3D Studio Max, the character's face is represented by a UV texture, which forms the basis for our face manipulation process [41]. For face manipulation, we employed an open-source implementation of 3DDFA_V2 [42]. 3DDFA_V2 is a deep learning-based algorithm that enables 3D dense face alignment. Its regression network framework strikes a balance between speed, accuracy, and stability. Fig. 3 illustrates the architecture of 3DDFA_V2, which comprises four main components: the lightweight backbone, landmark-regression regularization, meta-joint optimization of fWPDC and VDC [38], [39], and 3D-aided short-video synthesis.

In the 3DDFA_V2 architecture, input images are processed by the backbone network to extract features. These features are then passed through a pooling operation, generating two sets of tensors. One set of tensors is utilized to compute the landmark regression loss, while the other set is employed to calculate the Vertex Distance Cost (VDC) and Weighted Parameter Distance Cost (WPDC) [38], [39]. The meta-joint

TABLE I
THE ASSITIVE PROMPTS IN HIDE AND SEEK GAME.

Intensity Level	Assitive Prompts
1	Voice prompts for avatar's location
1	The avatar will wave at the user
2	A cartoon character appears to help the user using voice and gesture prompts
2	A bold red arrow appears to indicate the location of the avatar
3	The screen gradually darkens, and the visible area focuses on the avatar

Fig. 6. visualization of the assitive prompts in the hide and seek game.

optimization effectively combines the advantages of both cost functions. Additionally, the 3DDFA_V2 algorithm introduces a 3D-aided short-video synthesis method. By leveraging this technique, the network simulates both in-plane and out-of-plane facial movements, thereby transforming a still image into a short video. This approach enhances the stability of videos constructed from limited training data comprising only still images. Fig. 4 shows an example of face manipulation.

We now turn to the implementation details in the voice manipulation module. We invite a student speaker to record the voices of game dialog in an expressive manner, and collect 80 audio utterances from the user's parents. In terms of the voice conversion algorithm, we adopt the transformer [43] based Speaker Information Guided Voice Conversion system (SIG-VC) [40] to convert the prerecorded expressive dialog voice into each parent's characteristics. As described in Fig. 5, the SIG-VC system employs a supervised intermediate representation to reduce source speaker information and obtain purer content representations. SIG-VC uses a pre-trained speaker verification system for speaker representation extraction, and applys a pre-trained acoustic model [44] to extract the linguistic feature. Moreover, this system enforces intermediate representations in the vector space of selected acoustic features by sharing parameters for certain modules. Finally, a feedback loss [45], [46] is applied to assign high spoofing capability to the generated speech. The SIG-VC demo[1] is available online for readers.

2) Prompt Controller: The Prompt Controller plays a crucial role in assisting users who have difficulty finding the parent avatar. An effective instructional strategy for prompting is the least-to-most (LTM) prompting mechanism [47]. The LTM mechanism has been widely utilized in previous research to teach various skills, including motor skills [48] and communication skills [49], [50]. The LTM prompt mechanism employs a sequence of prompts to support the user during the game. When the Prompt Controller delivers an instruction, it utilizes a series of prompts, beginning with the least intrusive and progressively advancing to more instructive prompts. Prompts are presented individually until the user successfully finds the parent avatar. Typically, the LTM prompting mechanism employs a set of three consecutive prompts to facilitate the acquisition of a new skill. These prompts can include verbal cues, gestures, modeling, or physical assistance. In our system, the LTM mechanism allows the user to make their best effort to locate the hidden parent avatar before additional prompts are provided.

Table I and Fig. 6 show the assistive prompts in the hide-and-seek game. The level indicates the strength of the prompt. The larger the level, the more help the player gets. Fig. 7 shows the flow chart of the assistive prompt.

C. Gaze Module

In this study, we have developed a Gaze Module that utilizes the built-in 7Invensun[2] eye tracker to analyze gaze behavior in real-time. The eye tracker has a sampling frequency of 30 Hz and is integrated with the 7Invensun Unity development package. We collect continuous gaze coordinate data during the gameplay. Specifically, when the user's gaze remains within the avatar's body box in Unity for a duration of 500ms, the system determines that the player has successfully located the parent avatar and proceeds to the next round of the game. Our main focus in this study is on fixation data. To facilitate data collection, we have predefined multiple regions of interest (ROIs) within Unity. Whenever the user's gaze point falls within these predefined ROIs, the Gaze Module records and sends the corresponding gaze data to the Data Logger. As shown in Fig. 9, the ROIs are categorized into three types: face ROI, body ROI, and background ROI. The face ROI is defined by the avatar's facial region, which contains significant social and emotional information [7], [51]. Human face is a rich source of social information, and atypical gaze patterns on specific facial features are well-documented in ASD [52]. Due to the relatively long distance in the game scene, limitations in our hardware prevent the precise differentiation of fixation points on various parts of the face. The body ROI encompasses the avatar's body, providing cues that participants are still attending to the parent avatar, albeit with fewer social cues than the face region. The background ROI represents various

[1]https://haydencaffrey.github.io/sigvc/index.html
[2]https://www.7invensun.com/

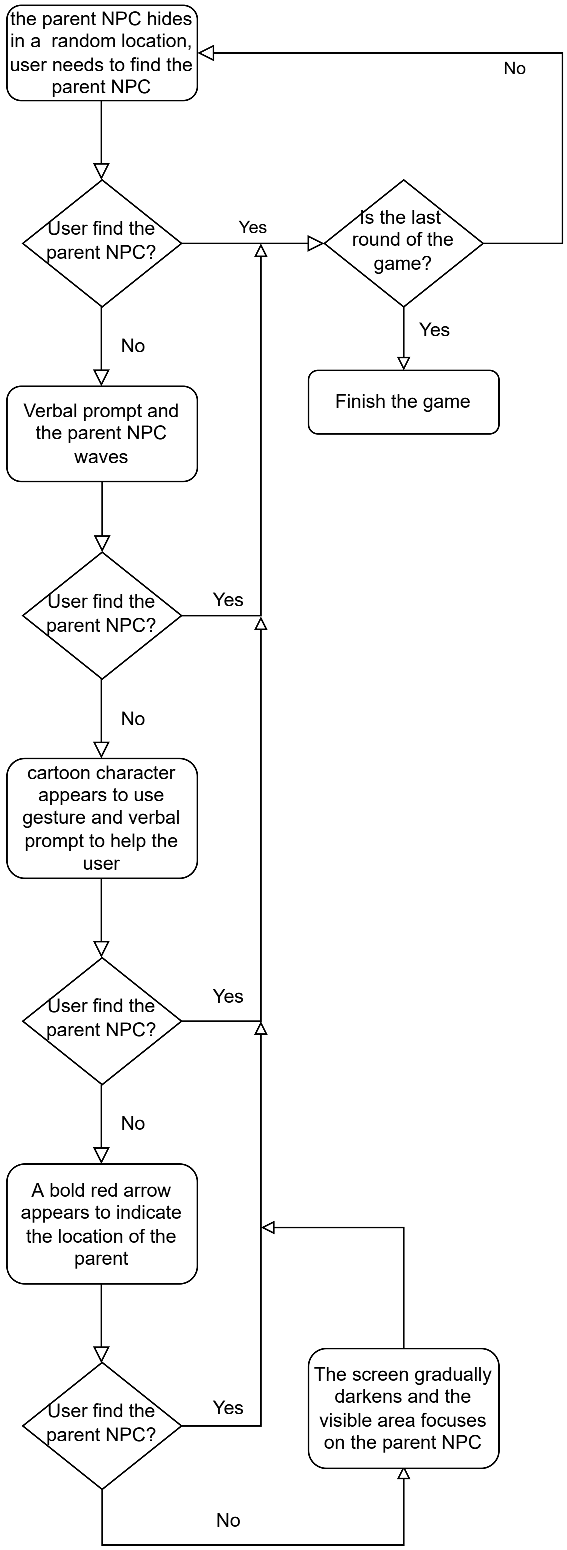


Fig. 7. Assistive flow chart in our prompt Controller. In the HSVRS system, we provide verbal, gestural and physical prompts. The prompts work together to guide participants to find the hidden parent avatar.

Fig. 8. The VR headset with its bulit-in gaze tracker.

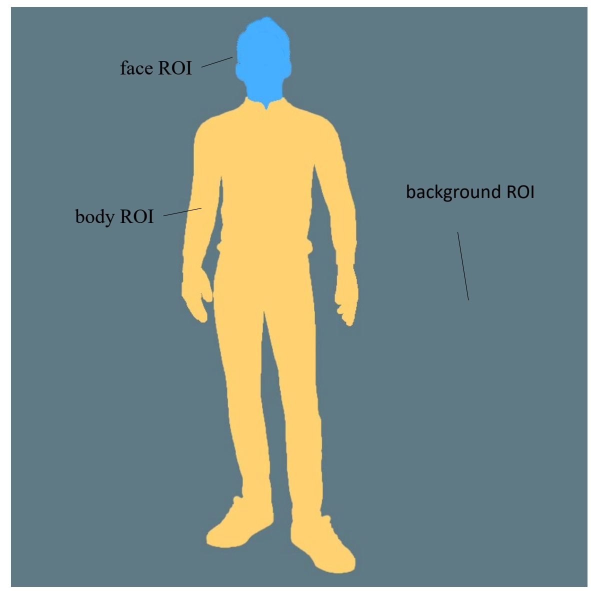


Fig. 9. The predefined three types of ROI: face ROI, body ROI and background ROI. Face ROI and body ROI are defined by avatar's face and body regions. The background ROI are defined by the scene's background.

elements in the virtual scene, such as furniture, walls, and other objects, which offer minimal social cues. During the gameplay, the Gaze Module captures the user's real-time gaze fixation data and transmits it to the Data Logger. We record the gaze fixation points and analyze the frequency of fixations across various ROIs. Once the game is completed, the Gaze Module calculates the proportion of fixation points within each respective ROI, and forwards these information to the Data Logger for further analysis.

D. Reinforcement Module

Reinforcement plays a crucial role in intervention systems, and it holds particular importance for children with autism [53], [54]. An effective reinforcement system can greatly enhance the effectiveness of an intervention [55], [56]. In the Hide and Seek Virtual Reality System (HSVRS), we have designed a comprehensive reinforcement system that offers various types of reinforcers to motivate and reward the children. Table II outlines the reinforcers available in the HSVRS system. The reinforcement system in HSVRS includes a wide range of visual and audio rewards. During the hide-and-seek game, when the child successfully finds the hiding location of the parent avatar, the Reinforcement Module provides verbal praise and thumbs-up gestures from the parent

TABLE II
THE REINFORCERS IN OUR HSVRS SYSTEM.

No.	Reinforcer	When the reinforcement arises
1	verbal praise of avatar characters	The user finds the parent avatar
2	thumbs up from the parent avatar	The user finds the parent avatar
3	The parent avatar performs a short dance	The user finds the parent avatar
4	display random cool effects like fireworks exploding	The user finds the parent avatar

Fig. 10. An example of reinforcer in our HSVRS system.

avatar as a form of positive reinforcement. Furthermore, in addition to the parent avatars, the virtual scenes are enriched with captivating special effects which are shown in Fig. 10. To further engage the children, HSVRS features a collection of pre-designed cartoon characters from which the children can choose their favorite character before beginning the game. As the hide-and-seek game progresses and all the hiding locations are found, the Reinforcement Module displays the child's selected cartoon character. This character will offer words of encouragement and perform an extended dance routine, providing additional visual and audio reinforcement.

E. Data Logger

The data logger collects data in virtual environment, including real-time gaze data and time stamps of different events during the game.

III. EXPERIMENTAL DESIGN

We conducted a pilot study to evaluate the hypothesis that incorporating the Hide and Seek Virtual Reality System (HSVRS) as an ancillary intervention would lead to improvements in gaze fixation among children with autism. This hypothesis was assessed through subjective questionnaire analysis. Furthermore, we aimed to compare the intervention effects between two groups: the avatar customized group and the uncustomized group. We hypothesized that children from the avatar customized group would demonstrate superior performance compared to those from the uncustomized group, as evidenced by reduced completion time and average response time, as well as an increased game score. Additionally, we explored whether the inclusion of face and voice deepfake manipulation had an impact on the gaze patterns observed across different Regions of Interest (ROIs). Before the start of experiment, the participants have a session to familiarize themselves with the VR headset through activities, e.g., touching the device and watching animations. Experiments start only after making sure that the child will not overreact to or reject the VR devices, and will immediately stop once the child feels discomfort. The duration of wearing VR devices was limited to ten minutes per session and each child is expected to join one session per day in a continuous 5-days period (T0-T4). Additionally, there are no breaks during a single session of experiment. All our experiments are conducted with explicit parental consent and active parental participation.

A. Participants

We recruited a total of thirty-six children with Autism Spectrum Disorder (ASD) from the Child Developmental & Behavioural Center of the Third Affiliated Hospital of Sun Yat-sen University in Guangzhou. We established three groups, including Avatar customized group, Avatar uncustomized group and control group. The control group only received normal artificial intervention while the Avatar customized group and the Avatar uncustomized group received VR-assisted interventions additional to normal artificial intervention. During the VR-assisted intervention, uniform avatars were used in the Avatar uncustomized group, while the avatars were customized referring to the participants' parents in the Avatar customized group. The artificial intervention primarily offered interventions within the natural context of structured Applied Behavior Analysis (ABA). The ABA intervention program used the ABC model to analyze behavioridentifying antecedents (what happens before the behavior), the behavior itself, and consequences (what happens after the behavior). Skills were broken down into smaller steps to make learning more manageable. Each step was taught systematically, and mastery of each step was required before moving on to the next. Initially, prompts were provided to help individuals perform a behavior. As the individual becomed more proficient, prompts were gradually decreased to promote independent functioning. ABA heavily relies on positive reinforcement to strengthen desired behaviors. Reinforcement was used to increase the likelihood of a behavior occurring in the future. Moreover, the parents of ASD children actively participated in the intervention process. They were trained with the core techniques of autism intervention.

B. Protocol

In this study, we adopt a subjective questionnaire to the parents of all participants both before and after the experiment. Additionally, we conducted five repeated trials over a session of seven days for both the avatar uncustomized and avatar customized groups. During each visit, children from both groups were instructed to play a game of hide-and-seek using our HSVRS system. It is important to note that the hiding location for each round of the game was randomly determined.

TABLE III
COMPARISON IN CLINICAL BASELINE DATA AMONG THE THREE GROUPS

	Control group	Avatar Uncustomized Group	Avatar Customized Group	F/Z/$\mathcal{X}^2$	P
Age	3.58(1.44)	5.25(1.60)	3.33(1.15)	11.39⋆	0.003
Gender				2.18†	0.336
male	12(100%)	10(83.33%)	11(91.67%)		
female	0(0.00%)	2(16.67%)	1(8.33%)		
ADOS					
Css	7.00(1.41)	7.71(1.70)	7.83(1.34)	1.05	0.361
FSIQ	73.50(19.92)	83.67(19.51)	70.25(14.12)	0.69	0.509
ABAS-II					
Overall adaptive skills	72.58(17.31)	83.67(19.51)	70.25(14.12)	2.10	0.138
Conceptual skills	84.00(17.62)	84.00(17.62)	75.00(13.21)	1.28	0.292
Social skills	71.33(18.57)	77.50(18.01)	65.42(18.08)	1.32	0.281
Practical skills	73.75(16.95)	87.92(20.92)	69.83(15.97)	3.32	0.048

[1] ADOS-Css: Autism Diagnostic Observation Schedule-Calibrated severity score; ABAS-II: Adaptive Behavior Assessment System Version II; FSIQ: full scale intelligence quotient
[2] † Chi-square test; ⋆ Kruskal-Wallis test; the others were One-Way Analysis of Variance

TABLE IV
PARENTS' FEEDBACK FROM PRE-TEST AND POST-TEST

Question	Pre-test						Post-test					
	ACG(n=12)		AUG(n=12)		CG(n=12)		ACG(n=12)		AUG(n=12)		CG(n=12)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
How often does the child perceive your gaze?	2.67	1.03	3.00	1.15	3.08	1.18	4.00	0.71	3.83	0.99	3.08	1.11
Can your child effectively respond to your pointing gesture?	2.92	0.95	3.08	0.76	3.83	0.80	4.42	0.86	4.08	0.86	3.92	0.95
Can your child effectively respond to your pointing gesture?	3.83	1.28	4.00	1.08	3.67	0.94	4.83	1.21	4.33	0.75	3.33	0.74
How often does your child engage in pointing?	3.67	0.62	4.33	1.18	3.75	0.83	4.83	1.21	4.58	1.14	3.33	1.03
How often does your child respond to auditory stimuli?	3.33	0.85	3.83	1.14	3.67	0.85	4.17	0.80	4.25	1.23	3.42	1.19
How often does your child respond to auditory stimuli?	3.33	1.60	3.92	1.32	3.83	0.99	4.50	1.12	4.42	1.11	3.67	0.94
Does the child enjoy playing hide-and-seek games?	4.17	1.21	4.25	1.36	3.75	1.23	4.83	0.90	4.75	1.30	3.42	1.19
Does the chid enjoy responding to verbal cues in hide-and-seek?	4.58	1.04	4.17	1.14	3.58	1.38	5.08	0.95	4.58	1.11	3.58	1.26
Does the child enjoy responding to body cues in hide-and-seek?	4.17	1.46	3.83	1.21	4.00	1.15	5.00	1.15	4.58	1.26	3.58	1.04
Does the child enjoy responding to cues from other people in hide-and-seek?	4.25	1.36	3.33	1.02	3.67	1.03	5.08	1.11	4.50	1.19	3.33	1.03

[1] CG: Control Group; AUG: Avatar Uncustomized Group; ACG: Avatar Customized Group

IV. RESULTS

In this section, we first collected baseline demographic information from all participants. Additionally, all participants with ASD were evaluated using the Autism Diagnostic Observation Schedule (ADOS [57]) and the Adaptive Behavior Assessment System, Second Edition (ABAS-II [58], [59]) during the baseline assessment. The baseline demographic information was analyzed to ensure that there were no statistically significant differences in the current levels of ASD symptoms among the participants in different groups.

Next, we analyzed the subjective questionnaires completed by the parents of the participants. The results of the questionnaire analysis are presented to demonstrate the positive effects of using our HSVRS as an auxiliary intervention therapy.

Furthermore, we present the objective results from two aspects: the player's game performance and the player's gaze fixation. We use mixed linear effect model to explore the effectiveness of the intervention. A mixed linear model (MLM), also known as a mixed-effects model or hierarchical linear model, is a statistical model that combines both fixed effects and random effects. It is an extension of the linear regression model, but it allows for the modeling of both fixed factors (predictors with fixed and known coefficients) and random factors (predictors with random and unknown coefficients). The longitudinal nature of our study requires a statistical approach to model and analyze the patterns of individual responses in the intervention session. The repeated measures structure of our data, wherein participants are measureed at multiple time points, requires a statistical model capable of handling correlated observations. The mixed linear model, by incorporating random effects, inherently addresses the issue of within-subject correlations. This is crucial in our study, where each participant undergoes multiple assessments over the course of the intervention. By comparing these results, we conclude that: i) our HSVRS is effective in improving children's gaze fixation ability, and ii) the use of deep learning-based face and voice manipulation technology can enhance the effectiveness of the intervention.

In the analysis of the experimental results, we utilized SPSS 26.0 software. Continuous data were presented as means and standard deviations (SD). A significance level of $P < 0.05$ was considered statistically significant.

A. Comparison in clinical baseline data

There is a significant difference in age among the three groups ($Z = 11.39, P = 0.003$), while no statistical difference in sex ratio ($\mathcal{X}^2 = 2.18, P = 0.336$). In order to reduce the difference in language levels and age, we applied standardized ADOS-Calibrated Severity Scores (Css) to evaluate the severe symptoms of autism. The results showed no statistical difference in ADOS-Css among the three groups ($F = 1.05, P = 0.361$), suggesting no significant difference in the severe symptoms of autism among the three groups. Similarly, standardized scores were applied to calculate age-normalized Intelligence quotient (IQ) and adaptive skills. There are no statistical differences in IQ and overall adaptive skills among the three groups (all $P > 0.05$), which indicated no significant difference in cognitive functioning and adaptive functioning among the three groups.

B. Analysis of subjective questionnaire

Before and after the experimental session, we collected parental feedback from the participants using a subjective questionnaire consisting of ten questions. The questions, along with the parents' responses from the pre-test and post-test, are presented in Table IV. A six-point Likert scale was used for all the questions. The questionnaire was divided into two parts, with the first six questions relating to a life scenario and the last four questions pertaining to the context of a hide-and-seek game. These questions assessed the child's ability to make eye contact, point with their finger, and engage in visual and auditory reasoning, all of which are skills addressed in our game. The total score of the questionnaire provided feedback on the child's overall ability. To compare the baseline data between the three groups, we conducted student t-tests on the total questionnaire scores. The results indicated no significant differences in the total questionnaire scores between the three groups (all $P > 0.05$).

We then divide the collected questionnaires into three groups: the control group, the avatar uncustomized group and the avatar customized group. Considering the repeated collection of the questionnaire, we conduct three mixed linear models with two components: 1) a random-effect parameter for each participant to account for unobserved sources of heterogeneity among participants, and 2) two fixed-effect parameters corresponding to the group and time.

Table V compares the difference in questionnaire points over time between the control group and the avatar uncustomized group. The score increased significantly between T4 and T0 in both groups ($P < 0.001$). Moreover, there are significant increment differences in questionnaire scores between the control group and the avatar uncustomized group. The increments of questionnaire scores were significantly larger in the avatar uncustomized group than in the control group. ($B = 0.61; P = 0.010$).

The change in questionnaire points over time between the control group and the avatar customized group is compared in Table VI between the two groups. Between T4 and T0, both groups' scores rose significantly ($P < 0.001$). Additionally, there are significant increases in questionnaire score discrepancies between the avatar customized group and the group that received no supplemental intervention. The increments in questionnaire scores in the avatar customized group were substantially more prominent than in the group that did not receive any auxiliary interventions ($B = 1.67; P < 0.001$).

Table VII compares the difference in questionnaire points over time between the avatar uncustomized group and the avatar customized group. The score increased significantly between T4 and T0 in both groups ($P < 0.001$). And there are significant increment differences in questionnaire scores between the avatar uncustomized group and the avatar customized group. The increments of questionnaire scores were significantly more prominent in the avatar customized group than in the avatar uncustomized group ($B = 0.61; P = 0.021$).

In summary, the VR auxiliary intervention group was more effective than the control group. Moreover, the best intervention results were achieved if avatars within the VR game were customized using face and voice deepfake manipulation techniques.

C. Analysis of participants' game performance

Game performance is measured using game scores, completion times, and average response times. Table VIII lists the metrics together with a description of each metric. Fig. 11 shows the line charts of the changes in the three game performance metrics over the intervention session. Participants' game performance, including completion time, average response time and game score, increase significantly between T4 and T0 in both groups. Specifically, children in the avatar

TABLE V
MIXED LINEAR MODEL IN COMPARING THE DIFFERENCE IN QUESTIONNAIRE OVER TIME BETWEEN THE CONTROL GROUP AND THE AVATAR UNCUSTOMIZED GROUP

Dependent variables	Intercept		Group effect		Time effect		Time-by-group effect	
	B (SE)	P value	B (SE)	P value	B (SE)	P value	B (SE)	P value
Questionnaire Points								
Baseline	36.72(2.30)	< 0.001	2.63(3.25)	0.425				
T0-T4					-1.02(0.54)	0.15	0.61(0.22)	0.010*

TABLE VI
MIXED LINEAR MODEL IN COMPARING THE DIFFERENCE IN QUESTIONNAIRE OVER TIME BETWEEN THE CONTROL GROUP AND THE AVATAR CUSTOMIZED GROUP

Dependent variables	Intercept		Group effect		Time effect		Time-by-group effect	
	B (SE)	P value	B (SE)	P value	B (SE)	P value	B (SE)	P value
Questionnaire Points								
Baseline	35.27(2.27)	< 0.001	1.19(3.22)	0.714				
T0-T4					-1.63(0.15)	0.001	1.67(0.77)	< 0.001***

TABLE VII
MIXED LINEAR MODEL IN COMPARING THE DIFFERENCE IN QUESTIONNAIRE OVER TIME BETWEEN THE AVATAR UNCUSTOMIZED GROUP AND THE AVATAR CUSTOMIZED GROUP

Dependent variables	Intercept		Group effect		Time effect		Time-by-group effect	
	B (SE)	P value	B (SE)	P value	B (SE)	P value	B (SE)	P value
Questionnaire Points								
Baseline	35.24(2.45)	< 0.001	-1.44(3.47)	0.416				
T0-T4					-1.64(0.17)	0.001	0.61(0.25)	0.021*

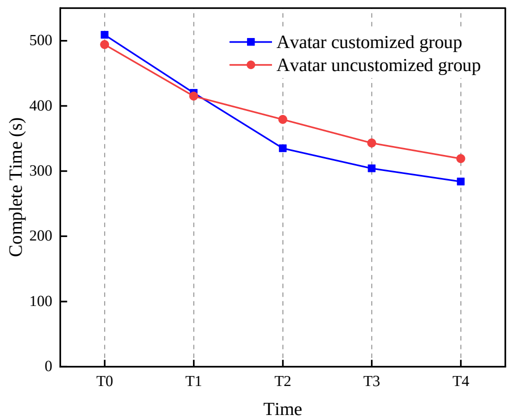
TABLE VIII
LIST OF PERFORMANCE METRICS

Performance Metric	Description
completion time (seconds)	The total time it takes for a participant to complete the game. In each game, the parent avatar hides in nine different random places. When the participant finds all nine places, the game is terminated.
Average Response Time (seconds)	Response time is computed between the time when the parent avatar hides properly and the time when the participant finds the hidden parent avatar.
Game Score (points)	In each round of the game, five points are received if the participant finds the parent avatar without any prompts. The game score decreases as the intensity of the prompt participant get increases. The maximum possible score is 45.

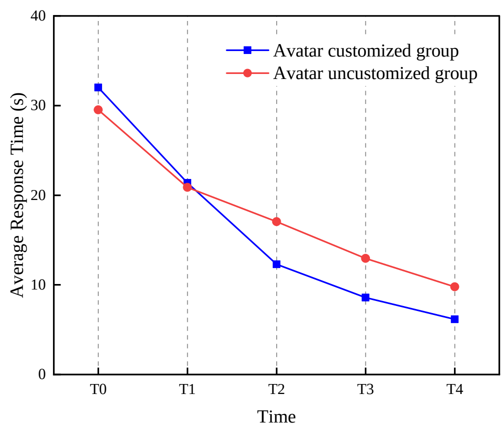
customized group took less time to complete the game, their average response time decreased, and they received higher game scores. For the completion time metric, the avatar uncustomized group and the avatar customized group had similar levels at T0, where they took about 500 seconds to complete the game. However, from Fig. 11, we find that children from the avatar customized group exhibited more significant improvement since the curve from the avatar customized group had a higher slope. For the average response time metric, the avatar uncustomized group and the avatar customized group showed similar levels at T0, where it takes roughly 32 seconds to discover the parent avatar in each round. Then, Fig. 11 shows that the average response time of the avatar-customized group reduces more quickly, suggesting that the children in this group received a better boost. For the game score metric, the average game score of the avatar customized group was lower than that of the uncustomized group at T0. Nevertheless, at T4, the average game score of the avatar customized group was higher than that of the uncustomized group. It indicates that children from the avatar customized group exhibited more remarkable improvement. [3]

Table IX shows the baseline assessment of the three game performance metrics. No statistical differences were found in completion time, average response time and game score between the avatar uncustomized group and the avatar customized group (all $P > 0.05$). It demonstrates that there is no

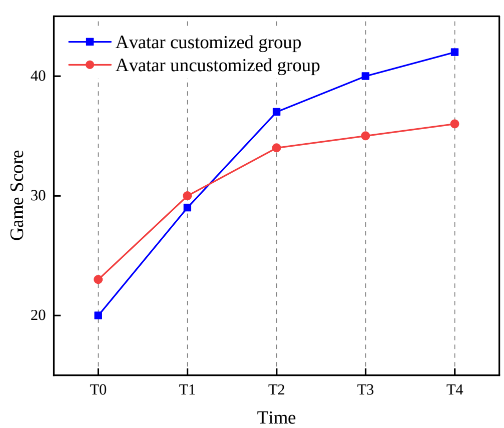
[3] A p-value below 0.05 receives a $*$, below 0.01 is marked with $**$, and below 0.001 is denoted with $***$.



(a) Time to Complete



(b) Average Response Time



(c) Game Score

Fig. 11. Line charts of the changes in the three game performance metrics over the intervention session.

significant difference between the avatar uncustomized group and the avatar customized group when children get started to play hide-and-seek in HRVRS.

For more detailed and convincing data analysis, as well as considering the repeated measures design of our study, a mixed linear model was applied to analyze outcome variables by modeling all five measurements (T0, T1, T2, T3 and T4).

Completion time, average response time and the game score increased significantly between T4 and T0 in both groups (all $P < 0.05$). After one week's auxiliary intervention, the increments in average response time and game score metrics were significantly larger in the avatar customized group than in the avatar uncustomized group (B=-2.28,1.99, P=0.001,0.001), whereas no significant differences in the increment scores were found in completion time metric(P=0.061). Thus, in summary, participants from the avatar customized group were able to achieve better intervention results.

D. Gaze Fixation Analysis

To better understand the distribution of the participants' fixation during the hide-and-seek game, we predefined three ROIs (regions of interest), namely face region, body region and background region as shown in Fig. 9. To obtain fixation metrics for these ROIs, we design an algorithm based on the 7Invensun Unity development package. For details, we record the fixation points of the gaze when the participants are playing the game, and then we divide the total fixation points into three categories based on their location. In the data analysis, we normalize these three metrics. The normalized results represent the fraction of fixation points from each ROI. From Table XI, we can observe that there is a statistically significant difference in face fixation proportion and background fixation proportions between the avatar uncustomized group and the avatar customized group (all $P < 0.001$). Meanwhile, there is no statistical difference in body fixation proportion between the avatar uncustomized group and the avatar customized group ($P = 0.141$). It demonstrate that using face and voice manipulation techniques to customize the characters in games made participants more likely to look at the facial regions of the avatars.

V. Discussion and Conclusion

In this study, we develop a novel VR system called HSVRS to target gaze fixation skills in autistic children. The system incorporates various components to enhance the intervention's effectiveness. The Avatar Controller enables avatar customization using computer vision and voice conversion techniques. The Prompt Controller accommodates diverse learning abilities by providing sequential prompts. The Game Module coordinates the game operation, while the Reinforcement Module provides rewards and positive feedback. The Gaze Module utilizes eye tracking for real-time analysis, and the Data Logger records all game and gaze data for analysis.

The pilot study aims to evaluate the efficacy of hide-and-seek games in VR headsets, specifically focusing on improving gaze fixation in children with ASD. We hypothesize that using HSVRS would lead to improved gaze fixation and that the

TABLE IX
COMPARISON IN COMPLETION TIME, AVERAGE RESPONSE TIME AND GAME SOCRE METRIC BETWEEN THE AVATAR UNCUSTOMIZED GROUP AND THE AVATAR CUSTOMIZED GROUP IN BASELINE (T0) ASSESSMENT

Dependent variables	avatar uncustomized group	avatar customized group	t value	P value
Game Performance				
completion time	495.58 (64.39)	466.92 (94.75)	0.867	0.288
Average Response Time	27.25 (11.99)	32.84 (7.48)	1.368	0.263
Game Score	23.08 (7.54)	26.33 (8.35)	-1.001	0.859

TABLE X
MIXED LINEAR MODEL IN COMPARING THE DIFFERENCE IN GAME PERFORMANCE OVER TIME BETWEEN THE AVATAR UNCUSTOMIZED GROUP AND THE AVATAR CUSTOMIZED GROUP

Dependent variables	Intercept		Group effect		Time effect		Time-by-group effect	
	B (SE)	P value	B (SE)	P value	B (SE)	P value	B (SE)	P value
completion time								
Baseline	526.51(20.83)	< 0.001	-47.72(29.47)	0.288				
T0-T4					-56.07(4.17)	< 0.001	-15.37(5.89)	0.061
Average response time								
Baseline	36.71(2.24)	< 0.001	-7.55(3.17)	0.263				
T0-T4					-6.53(0.46)	< 0.001	-2.28(0.64)	0.001**
Game Score								
Baseline	20.32(1.99)	< 0.001	5.37(2.82)	0.859				
T0-T4					4.59(0.42)	< 0.001	1.99(0.59)	0.001**

TABLE XI
RESULTS FOR GAZE FIXATIONS

Metrics	avatar un-customized group	avatar customized group		
	Mean(SD)	Mean(SD)	t value	P value
Face Fixation Propotion	12.99(2.87)	22.04(3.94)	6.43	< 0.001***
Body Fixation Propotion	21.72(5.46)	21.54(2.25)	-1.08	0.141
Background Fixation Propotion	65.28(6.88)	56.41(4.99)	-3.613	< 0.001***

avatar customized group would exhibit larger increments than the uncustomized avatar group. The participants are divided into three groups, and repeated measurements are conducted over a one-week session. The experimental results demonstrate the potential of the HSVRS system to improve gaze fixation in children with ASD. Both groups show improvements in game performance, as evidenced by shorter completion times, shorter average response times, and higher game scores. Furthermore, the mixed linear model analysis reveals significant and consistent inter-group differences. After one week of intervention, the avatar customized group exhibits significantly larger increments in average response time and game score compared to the avatar uncustomized group. These findings support the hypothesis that in-game avatar customization enhances intervention effectiveness. The gaze data analysis reveals that children from the avatar customized group focus more on the face region of the parent avatar, indicating the positive impact of face manipulation design in promoting eye contact.

The transferability of skills and experiences from VR to real-life contexts is an important aspect of any virtual intervention. Our HSVRS, with its animated environment, is designed not only to engage children but also to simulate real-world social interactions and gaze patterns. The pre-test and post-test subjective questionnaire partially show the effectiveness of real-life gaze fixation ability improvement. For example, results in Table IV indicate that children participated in the VR-assisted intervention showed significant improvements in gaze fixation abilities. These improvements suggest that the skills learned in the VR environment may generalize to real-world interactions. However, we need to admit the limitation that more comprehensive pre- and post-test professional behavior assessment should have been performed to better validate the improvement and transferability of skills in real-life.

Our study introduces a novel approach to evaluate ASD treatments by utilizing gaze fixation data collected through an immersive virtual reality environment. This methodological innovation extends previous research by providing a feasible framework of how children with ASD engage with social cues in a controlled and interactive setting. The eye-tracking technology in VR provides a more accurate and detailed analysis of gaze patterns. Another innovation of our proposed method is the technique of deep learning-based face and voice manipulation not only makes the intervention more engaging but also provides a personalized learning experience tailored to the child's familiarity and comfort, which potentially could enhance the transfer of skills to real-world situations. From a practical viewpoint, the HSVRS offers a promising complement to traditional treatments like ABA, which is resource-intensive and accessible only to a limited number of families, particularly in regions with limited specialized services. Our proposed system is scalable and affordable for families and institutions. Compared with traditional treatment, our proposed system can be implemented in various environments ranging from specialized clinics to home and be accessible from a broader range of children with ASD.

VI. Future Work

While the results are encouraging, it is important to acknowledge the limitations of this study and identify areas for future investigation. The sample size is relatively small, and a larger longitudinal study would provide more convincing results. Currently, our questionnaire is mainly targeting the parents. To improve the reliability of our findings, we plan to incorporate surveys targeting ASD children, employing simpler and more comprehensible questionnaires. Additionally, future studies should include pre- and post-test behavior assessment to further evaluate the impact of HSVRS on gaze fixation ability in real-life. Overall, this pilot study demonstrates the potential benefits of using HSVRS for improving gaze fixation in children with ASD, providing a foundation for future research in this field.

Building upon the findings of this study, we will start a more comprehensive and longitudinal research initiative that will pay attention to the long-term effects and potential therapeutic benefits of our proposed VR games. The following outlines our proposed future works:

- Longitudinal Study Design: We plan to conduct a longitudinal study that spans several months, during which participants will engage with the VR games over multiple sessions. The study will include pre-intervention, post-intervention, and follow-up assessments at regular intervals to evaluate the sustainability of the intervention effects.
- Measuring Long-Term Impact: We will employ standardized assessments and questionnaires at each follow-up to measure the long-term impact on the children's social skills and overall quality of life. The assessment will include both quantitative and qualitative measures to capture a comprehensive view of the children's progress.
- Lighter device and more games: We plan to extend the VR games from one to four, including finding similarities, blowing bubbles, bamboo dragonflies, these three new games. Also, devices with lighter weight will be used to improve wearing comfort and extend the duration of the experiment.

References

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The lancet*, vol. 392, no. 10146, pp. 508–520, 2018.

[2] A. P. Association, A. P. Association *et al.*, "Dsm-5 task force diagnostic and statistical manual of mental disorders: Dsm-5. washington," *DC: American Psychiatric Association*, 2013.

[3] S. Wallace, D. Fein, M. Rosanoff, G. Dawson, S. Hossain, L. Brennan, A. Como, and A. Shih, "A global public health strategy for autism spectrum disorders," *Autism Research*, vol. 5, no. 3, pp. 211–217, 2012.

[4] M. J. Maenner, K. A. Shaw, A. V. Bakian, D. A. Bilder, M. S. Durkin, A. Esler, S. M. Furnier, L. Hallas, J. Hall-Lande, A. Hudson *et al.*, "Prevalence and characteristics of autism spectrum disorder among children aged 8 yearsautism and developmental disabilities monitoring network, 11 sites, united states, 2018," *MMWR Surveillance Summaries*, vol. 70, no. 11, p. 1, 2021.

[5] S. J. Webb, E. J. Jones, J. Kelly, and G. Dawson, "The motivation for very early intervention for infants at high risk for autism spectrum disorders," *International journal of speech-language pathology*, vol. 16, no. 1, pp. 36–42, 2014.

[6] C. Chevallier, J. Parish-Morris, A. McVey, K. M. Rump, N. J. Sasson, J. D. Herrington, and R. T. Schultz, "Measuring social attention and motivation in autism spectrum disorder using eye-tracking: Stimulus type matters," *Autism Research*, vol. 8, no. 5, pp. 620–628, 2015.

[7] W. Jones and A. Klin, "Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, pp. 427–431, 2013.

[8] T. W. Frazier, M. Strauss, E. W. Klingemier, E. E. Zetzer, A. Y. Hardan, C. Eng, and E. A. Youngstrom, "A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 56, no. 7, pp. 546–555, 2017.

[9] J. O. Cooper, T. E. Heron, W. L. Heward *et al.*, "Applied behavior analysis," 2007.

[10] G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley, "Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model," *Pediatrics*, vol. 125, no. 1, pp. e17–e23, 2010.

[11] S. Carlon, M. Carter, and J. Stephenson, "A review of declared factors identified by parents of children with autism spectrum disorders (asd) in making intervention decisions," *Research in Autism Spectrum Disorders*, vol. 7, no. 2, pp. 369–381, 2013.

[12] C. Kasari and T. Smith, "Interventions in schools for children with autism spectrum disorder: Methods and recommendations," *Autism*, vol. 17, no. 3, pp. 254–267, 2013.

[13] M. M. Smith and I. Connolly, "Roles of aided communication: Perspectives of adults who use aac," *Disability and rehabilitation: Assistive technology*, vol. 3, no. 5, pp. 260–273, 2008.

[14] S. Bekteshi, M. Konings, I. Vanmechelen, J. Deklerck, E. Ortibus, J.-M. Aerts, H. Hallez, P. Karlsson, B. Dan, and E. Monbaliu, "Eye gaze gaming intervention in children with dyskinetic cerebral palsy: a pilot study of task performance and its relation with dystonia and choreoathetosis," *Developmental neurorehabilitation*, vol. 23, no. 8, pp. 548–556, 2020.

[15] S. Parsons and S. Cobb, "State-of-the-art of virtual reality technologies for children on the autism spectrum," *European Journal of Special Needs Education*, vol. 26, no. 3, pp. 355–366, 2011.

[16] B. G. Travers, A. H. Mason, L. A. Mrotek, A. Ellertson, D. C. Dean, C. Engel, A. Gomez, O. I. Dadalko, and K. McLaughlin, "Biofeedback-based, videogame balance training in autism," *Journal of autism and developmental disorders*, vol. 48, pp. 163–175, 2018.

[17] F. Berton, A.-H. Olivier, J. Bruneau, L. Hoyet, and J. Pettré, "Studying gaze behaviour during collision avoidance with a virtual walker: Influence of the virtual reality setup," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 717–725.

[18] F. Berton, L. Hoyet, A.-H. Olivier, J. Bruneau, O. Le Meur, and J. Pettré, "Eye-gaze activity in crowds: impact of virtual reality and density," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 322–331.

[19] V. Yaneva, S. Eraslan, Y. Yesilada, R. Mitkov *et al.*, "Detecting high-functioning autism in adults using eye tracking and machine learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 6, pp. 1254–1261, 2020.

[20] M. Courgeon, G. Rautureau, J.-C. Martin, and O. Grynszpan, "Joint attention simulation using eye-tracking and virtual humans," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 238–250, 2014.

[21] N. Caruana, H. Stieglitz Ham, J. Brock, A. Woolgar, N. Kloth, R. Palermo, and G. McArthur, "Joint attention difficulties in autistic adults: an interactive eye-tracking study," *Autism*, vol. 22, no. 4, pp. 502–512, 2018.

[22] G. Little, L. Bonnar, S. Kelly, K. S. Lohan, and G. Rajendran, "Gaze contingent joint attention with an avatar in children with and without asd," in *2016 joint IEEE international conference on development and learning and epigenetic robotics (ICDL-EpiRob)*. IEEE, 2016, pp. 15–20.

[23] K. Porayska-Pomsta, A. M. Alcorn, K. Avramides, S. Beale, S. Bernardini, M. E. Foster, C. Frauenberger, J. Good, K. Guldberg, W. Keay-Bright *et al.*, "Blending human and artificial intelligence to support autistic childrens social communication skills," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 25, no. 6, pp. 1–35, 2018.

[24] S. Bernardini, K. Porayska-Pomsta, and T. J. Smith, "Echoes: An intelligent serious game for fostering social communication in children with autism," *Information Sciences*, vol. 264, pp. 41–60, 2014.

[25] J. Zhao, X. Zhang, Y. Lu, X. Wu, F. Zhou, S. Yang, L. Wang, X. Wu, and F. Fei, "Virtual reality technology enhances the cognitive and social communication of children with autism spectrum disorder," *Frontiers in Public Health*, vol. 10, p. 1029392, 2022.

[26] J. Li, Z. Zheng, Y. Chai, X. Li, and X. Wei, "Faceme: An agent-based social game using augmented reality for the emotional development of children with autism spectrum disorder," *International Journal of Human-Computer Studies*, vol. 175, p. 103032, 2023.

[27] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, "Affective interactions using virtual reality: the link between presence and emotions," *Cyberpsychology & behavior*, vol. 10, no. 1, pp. 45–56, 2007.

[28] B. Karami, R. Koushki, F. Arabgol, M. Rahmani, and A.-H. Vahabie, "Effectiveness of virtual/augmented reality–based therapeutic interventions on individuals with autism spectrum disorder: a comprehensive meta-analysis," *Frontiers in Psychiatry*, vol. 12, p. 665326, 2021.

[29] H. Sandgreen, L. H. Frederiksen, and N. Bilenberg, "Digital interventions for autism spectrum disorder: a meta-analysis," *Journal of Autism and Developmental Disorders*, vol. 51, pp. 3138–3152, 2021.

[30] P. Mesa-Gresa, H. Gil-Gómez, J.-A. Lozano-Quilis, and J.-A. Gil-Gómez, "Effectiveness of virtual reality for children and adolescents with autism spectrum disorder: an evidence-based systematic review," *Sensors*, vol. 18, no. 8, p. 2486, 2018.

[31] H. H. Ip, S. W. Wong, D. F. Chan, J. Byrne, C. Li, V. S. Yuan, K. S. Lau, and J. Y. Wong, "Enhance emotional and social adaptation skills for children with autism spectrum disorder: A virtual reality enabled approach," *Computers & Education*, vol. 117, pp. 1–15, 2018.

[32] J. P. McCleery, A. Zitter, R. Solórzano, S. Turnacioglu, J. S. Miller, V. Ravindran, and J. Parish-Morris, "Safety and feasibility of an immersive virtual reality intervention program for teaching police interaction skills to adolescents and adults with autism," *Autism Research*, vol. 13, no. 8, pp. 1418–1424, 2020.

[33] B. J. Li, R. Ratan, and M. O. Lwin, "Virtual game changers: How avatars and virtual coaches influence exergame outcomes through enactive and vicarious learning," *Behaviour & Information Technology*, vol. 41, no. 7, pp. 1529–1543, 2022.

[34] N. Yee and J. Bailenson, "The proteus effect: The effect of transformed self-representation on behavior," *Human communication research*, vol. 33, no. 3, pp. 271–290, 2007.

[35] E. Bekele, J. A. Crittendon, A. Swanson, N. Sarkar, and Z. E. Warren, "Pilot clinical application of an adaptive robotic system for young children with autism," *Autism*, vol. 18, no. 5, pp. 598–608, 2014.

[36] U. Lahiri, E. Bekele, E. Dohrmann, Z. Warren, and N. Sarkar, "Design of a virtual reality based adaptive response technology for children with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 1, pp. 55–64, 2012.

[37] M. T. Brodhead, T. S. Higbee, J. S. Pollard, J. S. Akers, and K. R. Gerencser, "The use of linked activity schedules to teach children with autism to play hide-and-seek," *Journal of applied behavior analysis*, vol. 47, no. 3, pp. 645–650, 2014.

[38] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2016, pp. 146–155.

[39] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 78–92, 2017.

[40] H. Zhang, Z. Cai, X. Qin, and M. Li, "Sig-vc: A speaker information guided zero-shot voice conversion system for both human beings and machines," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6567–65 571.

[41] T. Mullen, *Mastering blender*. John Wiley & Sons, 2011.

[42] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 152–168.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[44] D. D. Deliyski, "Acoustic model and evaluation of pathological voice production," in *Third European Conference on Speech Communication and Technology*, 1993.

[45] Z. Cai, C. Zhang, and M. Li, "From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint," in *Proc. of Interspeech*, 2020, pp. 3974–3978.

[46] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing voice conversion network with cycle consistency loss of speaker identity," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 507–513.

[47] M. E. Libby, J. S. Weiss, S. Bancroft, and W. H. Ahearn, "A comparison of most-to-least and least-to-most prompting on the acquisition of solitary play skills," *Behavior analysis in practice*, vol. 1, no. 1, pp. 37–43, 2008.

[48] M. Yanardag, B. Birkan, İ. Yılmaz, F. K. Konukman, B. Ağbuğa, and L. Lieberman, "The effects of least-to-most prompting procedure in teaching basic tennis skills to children with autism," *Kinesiology*, vol. 43, no. 1., pp. 44–55, 2011.

[49] E. H. Finke, J. M. Davis, M. Benedict, L. Goga, J. Kelly, L. Palumbo, T. Peart, and S. Waters, "Effects of a least-to-most prompting procedure on multisymbol message production in children with autism spectrum disorder who use augmentative and alternative communication," *American journal of speech-language pathology*, vol. 26, no. 1, pp. 81–98, 2017.

[50] A. S. Polick, J. E. Carr, and N. M. Hanney, "A comparison of general and descriptive praise in teaching intraverbal behavior to children with autism," *Journal of Applied Behavior Analysis*, vol. 45, no. 3, pp. 593–599, 2012.

[51] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "Human neural systems for face recognition and social communication," *Biological psychiatry*, vol. 51, no. 1, pp. 59–67, 2002.

[52] M. Chita-Tegmark, "Social attention in asd: A review and meta-analysis of eye-tracking studies," *Research in developmental disabilities*, vol. 48, pp. 79–93, 2016.

[53] J. A. Williams, R. L. Koegel, and A. L. Egel, "Response-reinforcer relationships and improved learning in autistic children," *Journal of Applied Behavior Analysis*, vol. 14, no. 1, pp. 53–60, 1981.

[54] S. Kang, M. OReilly, L. Rojeski, K. Blenden, Z. Xu, T. Davis, J. Sigafoos, and G. Lancioni, "Effects of tangible and social reinforcers on skill acquisition, stereotyped behavior, and task engagement in three children with autism spectrum disorders," *Research in developmental disabilities*, vol. 34, no. 2, pp. 739–744, 2013.

[55] M. H. Charlop-Christy and L. K. Haymes, "Using objects of obsession as token reinforcers for children with autism," *Journal of Autism and Developmental Disorders*, vol. 28, no. 3, pp. 189–198, 1998.

[56] R. L. Koegel, T. W. Vernon, and L. K. Koegel, "Improving social initiations in young children with autism using reinforcers with embedded social interactions," *Journal of autism and developmental disorders*, vol. 39, no. 9, pp. 1240–1251, 2009.

[57] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[58] T. Oakland and P. L. Harrison, *Adaptive behavior assessment system-II: Clinical use and interpretation*. Academic Press, 2011.

[59] ——, "Adaptive behaviors and skills: An introduction," in *Adaptive Behavior Assessment System-II*. Academic Press, 2008, pp. 1–20.