Computer-Aided Autism Spectrum Disorder Diagnosis With Behavior Signal Processing

Ming Cheng, Yingying Zhang, Yixiang Xie, Yueran Pan, Xiao Li, Wenxing Liu, Chengyan Yu, Dong Zhang, Yu Xing, Xiaoqian Huang, Fang Wang, Cong You, Yuanyuan Zou, Yuchong Liu, Fengjing Liang, Huilin Zhu, Chun Tang, Hongzhu Deng, Xiaobing Zou, and Ming Li

***Abstract*—Behavioral observation plays an essential role in the diagnosis of Autism Spectrum Disorder (ASD) by analyzing children's atypical patterns in social activities (e.g., impaired social interaction, restricted interests, and repetitive behavior). To date, this process still heavily relies on the questionnaire survey, clinical observation, or retrospective video analysis, leading to high demand for professionals with massive labor costs. This article proposes a standardized platform for stimulating, gathering, analyzing, modeling, and interpreting human behavioral data in the application of computer-aided ASD diagnosis. By a structured assessment process, the proposed system can automatically evaluate children's multiple social interaction skills using the captured audio-visual data and provide the final diagnostic suggestions. We collect a multimodal behavioral database of 95 participants (71 children with ASD and 24 age-matched typical controls) in a real clinic environment, the Third Affiliated Hospital of Sun Yat-sen University, China. On the clinical database, our proposed computer-aided ASD diagnosis system obtains an accuracy of 88.42% for identifying ASD children with an average age of 24 months, representing a performance comparable to top-level human experts. As a unified and replicable solution, it has good potential to be promoted to less developed areas with limited high-quality medical resources.**

***Index Terms*—Computer-aided ASD diagnosis, multimodal behavior signal processing.**

Manuscript received 8 August 2022; revised 10 January 2023; accepted 18 January 2023. Date of publication 23 January 2023; date of current version 29 November 2023. This work was supported in part by the Science and Technology Program of Guangzhou under Grants 202007030011 and 201903010040, in part by the National Natural Science Foundation of China under Grants 62171207, 81873801, and 62173353, and in part by the Medical Science and Technology Research Foundation of Guangdong Province under Grant A2022039. *(Ming Cheng and Yingying Zhang contributed equally to this work.)* Recommended for acceptance by M. Mahoor. *(Corresponding authors: Xiaobing Zou and Ming Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the IRB committee of 3rd affiliated hospital of Sun Yat-sen University under application No. [2018] 02-196-01 and the IRB committee of Duke Kunshan University under application No. 2018LIM024 and performed in line with Multimodal human behavior modeling through computational sensing and analysis for young children with autism spectrum disorders.

Ming Cheng, Yueran Pan, Wenxing Liu, and Ming Li are with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with the Data Science Research Center, Duke Kunshan University, Kunshan 215316, China (e-mail: ming.cheng@whu.edu.cn; panyr.math@whu.edu.cn; wenxing.liu@dukekunshan.edu.cn; ming.li369@dukekunshan.edu.cn).

Yingying Zhang, Yu Xing, Xiaoqian Huang, Fang Wang, Cong You, Yuanyuan Zou, Yuchong Liu, Fengjing Liang, Huilin Zhu, Chun Tang, Hongzhu Deng, and Xiaobing Zou are with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China (e-mail: summer09251228@163.com; 583101992@qq.com; huangxq79@mail.sysu.edu.cn; wangfang6056@163.com; 474159517@qq.com; 519138879@qq.com; laouchong@mail.sysu.edu.cn; liangfj3@mail.sysu.edu.cn; zhuhlin6@mail.sysu.edu.cn; tch020@126.com; denghz@mail.sysu.edu.cn; zouxb@mail.sysu.edu.cn).

Yixiang Xie, Xiao Li, Chengyan Yu, and Dong Zhang are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: xieyx55@mail.sysu.edu.cn; lixiao37@mail2.sysu.edu.cn; yuchy33@mail2.sysu.edu.cn; zhangd@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TAFFC.2023.3238712

I. Introduction

AUTISM Spectrum Disorder (ASD) is a neurodevelopmental disorder with the core characteristic of defects in social communication [1]. According to the U.S. Centers for Disease Control and Prevention (CDC), the estimated ASD prevalence among 8-year-old American children increased to 2.3% in 2018 [2]. As a neurodiverse disorder, the prevalence of ASD will bring critical impacts on people's way of life and lead to high social welfare costs.

The intervention for autistic children should begin as early as possible [3]. However, the lack of experienced and professional experts limits the accessibility of early-stage ASD identification. For instance, Autism Diagnostic Observational Schedule (ADOS) [4] is a widely-used clinical approach, and a well-trained doctor needs to spend an average of 10 to 20 hours on each assessment case [5]. Moreover, the interpretation of diagnostic outcomes heavily depends on the doctor's expertise. Hence, promoting early diagnosis is difficult, especially in underdeveloped regions.

Many previous studies have exploited to enhance the efficiency of clinical methods by incorporating computer technologies. Some researchers build vision-based systems to model autistic children's abnormalities in social responses [6], head movements [7], gaze patterns during face-to-face conversations [8], and facial expressions [9]. Based on speech processing, acoustic features can also help distinguish the ASD group from the typical development [10], [11]. Furthermore, the wearable accelerator can help detect ASD children's stereotypical motor movement (SMM) [12], [13]. Nevertheless, there is no universal framework for stimulating, gathering, analyzing, modeling, and interpreting various behavioral data. Those approaches are still limited to laboratory settings.

The highlight of this paper is to propose a novel computer-aided ASD diagnosis system based on multimodal behavior signal processing, shown in Fig. 1. We first develop a specialized testing studio for social activities, which can present standardized and objective audiovisual stimuli and capture the

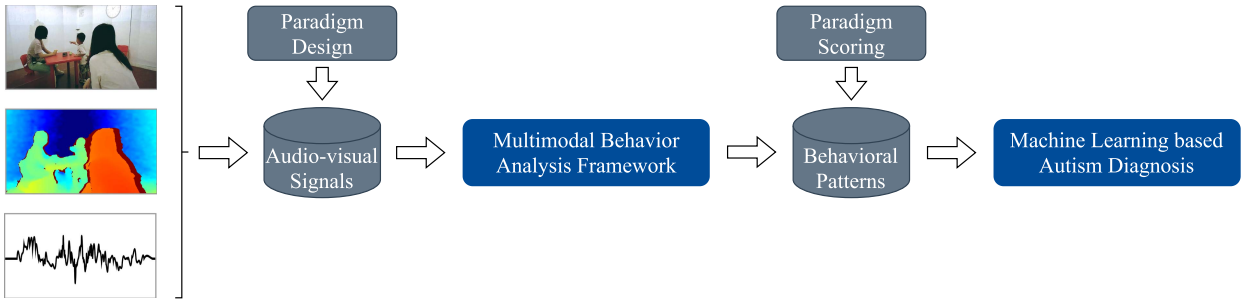


Fig. 1. System design of the proposed ASD diagnosis framework. The complete system comprises hardware integration, autism domain knowledge, and computer algorithms, supporting the automatic ASD diagnosis without the high demand for well-trained clinicians.

participants' multimodal behavioral data. Then, we introduce autism-related domain knowledge to build structured assessment paradigms for measuring children's social interaction skills in the testing studio. Furthermore, a multimodal behavior signal processing framework is adopted to recognize and score the participants' behavioral characteristics. By collecting 95 participants' assessment scores and behavior features, machine learning classifiers with leave-one-out cross-validation are trained to provide the final diagnostic suggestions. To our knowledge, this work is one of the first computer-aided ASD diagnosis systems tested in a realistic clinical environment. The contributions can be summarized as follows:

- We develop a *standardized and editable* platform for gathering and analyzing behavioral data during the assessment process. Doctors can set up a new stimulus in a few steps, making the platform more flexible in clinical research.
- We design a *structured* assessment process to evaluate children's social interaction skills by relatively comprehensive paradigms. The following ASD diagnosis is *interpretable* from the clinical point of view.
- The multimodal data acquisition and behavior signal processing are *unconstrained* in the testing studio. Participants can move freely in the testing studio. All social activities take place in natural interactions.
- Finally, our proposed computer-aided ASD diagnosis system obtains an accuracy of 88.42% on the collected database of 95 participants with an average age of 24 months, showing *accurate* performance comparable with human experts.

II. Related Works

A. Clinical Methods

Currently, clinical ASD screening and diagnosis mainly rely on two ways: investigation of medical history by scoring checklists and behavioral observation [3].

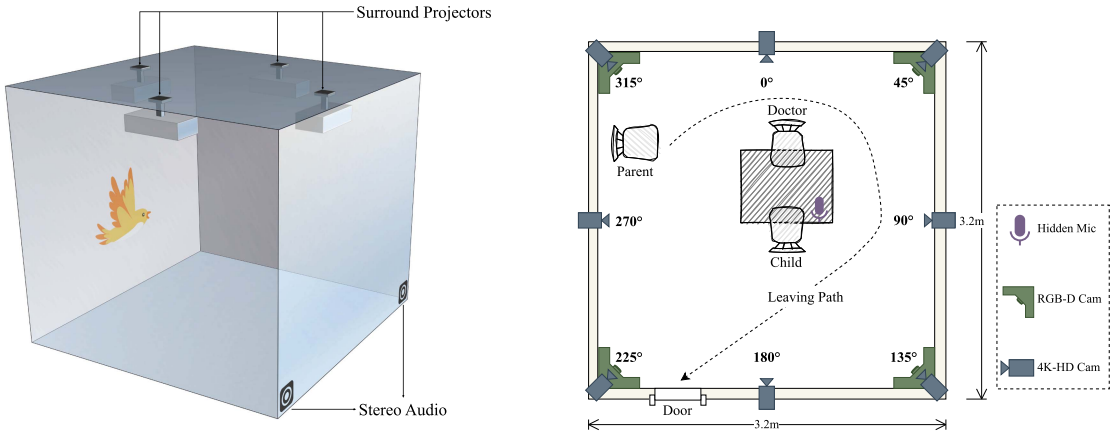
The Autism Diagnostic Interview - Revised (ADI-R) [14] is a questionnaire-based scoring checklist for rating autism risks. It contains a series of interview questions for parents to describe children's behavioral patterns in daily life. Then, doctors can provide the assessment results by corresponding evaluation metrics. Moreover, M-CHAT [15], M-CHAT-R/F [16], and Autism Behavior Checklist (ABC) [17] are also widely used scoring checklists in clinics.

Behavioral observation is another measure for autism screening and diagnosis. The Autism Diagnostic Observational Schedule (ADOS) [4] and its revised version, ADOS-2 [18], are representative instruments in this field. They consist of several coding components to evaluate children's behavioral performance in designed interactive sessions, which can provide a reference for the following clinical diagnosis. Usually, the worldwide gold standard of autism diagnosis is adopting the combination of ADOS [4] and ADI-R [14]. For children under 18 months, the doctor's clinical diagnosis is used as the primary suggestion [19].

B. Computer-Aided Methods

In recent years, how computer technologies can aid in ASD screening has become a popular topic. The most practical solution is to speed up the existing clinical diagnosis process through machine learning. Dennis Wall et al. [20] utilize the ADTree algorithm to investigate the effectiveness of assessment indicators in ADOS Module-1 [18]. The feature selection conducted on 612 participants shows that using only 8 of the 29 items can achieve comparable effects with full data. Subsequent studies [21] extend this work to the ADOS Module2 and Module-3 data within 4540 individuals, indicating that 9/28 items of Module-2 and 12/28 items of Module3 are adequate to detect ASD risks with the accuracy of 98.27% and 97.66%, respectively.

With deep learning (DL) breakthroughs, machines can help automatically detect and analyze behavioral signs related to ASD risks, e.g., atypical visual attention [22], [23], difficulties in orienting to name calls, social reference, and responsive social smile [24], [25], [26], [27]. Jordan Hashemi et al. [6] design a mobile application to engage children's attention and social responses, with cameras and microphones to capture and analyze the audio-visual data of ASD and Non-ASD children's behaviors. Rujing Zhang et al. [28] develop computer games and quantitative indicators to evaluate ASD children's visual perception, eye-hand coordination, and fine motor skills. Tanaya Guha et al. [9] utilize a computational method to figure out that the complexity of facial emotions in children with high-functioning autism (HFA) is less than in the regular group. Fenglei Zhu et al. [29] develop a multimodal perception system to quantify children's behaviors in Response-to-Name procedures. Moreover, the integrations of computer vision, wearable accelerators and neural networks are also explored to provide stereotypical motor movement reports in children's daily life [12], [13], [30].



(a) Multimedia Output (b) Sensor Layout

Fig. 2. Designed testing studio. The ceiling-mounted projectors and stereo loudspeakers can display standardized multimedia stimuli, providing a natural and immersive environment. For instance, when the child is playing with others, a bird may appear next to him or her to test the reactions to the environment. Meanwhile, comprehensive behavior data will be recorded synchronously into video, depth, and audio signals.

Also, DL-based methods can be used to output ASD diagnostic suggestions based on multimodal data directly. Usually, these models follow the way of pre-training on fundamental tasks with large-scale databases and finetuning on ASD-specific tasks to solve the problem of limited training data. Chin-Po Chen et al. [31] design several multimodal behavior descriptors to discriminate different ASD subgroups by analyzing their audio-visual data in ADOS-based social interactions. Hung-Yi Lee et al. [10] design a learnable acoustic segment model (ASM) to implement an ensemble system for identifying ASD children via voice data. Ming Li et al. [11] gather a speech corpus of mandarin communications recorded in the ADOS sessions, then propose an automated assessment framework to detect speech and language abnormalities in toddlers with ASD. The head movement [7], facial appearance [32], and gaze data [8] are also discriminative features in identifying ASD.

Moreover, biomedical testing technologies bring new opportunities and reveal more in-depth findings. The Autism Brain Imaging Data Exchange (ABIDE) datasets [33], [34] release over 1,000 magnetic resonance imaging (MRI) samples with ASD and associated symptoms, triggering multiple subsequent studies adopting MRI data and deep learning to the autism diagnosis [35], [36]. Furthermore, gene analysis [37] and electroencephalogram (EEG) signal processing [38], [39] have attracted a rising number of attempts in recent years.

The advantages and disadvantages of the aforementioned computer-aided methods can be summarized as follows. First, the methods based on clinical document data (e.g., ADOS, ADI-R) have superior validity and interpretability, while the prerequisite of clinical assessments highly depends on well-trained clinicians. Second, the methods based on data-driven behavior analysis usually can work with some low-cost sensors (e.g., cameras, microphones) without high demand for professionals. However, the lack of interpretability in the black-box diagnostic process is a big challenge for clinical practice. Third, the gene analysis, MRI, and EEG methods need high equipment costs. They are promising but still need to develop more.

To address these problems, we propose a computer-aided diagnosis system to identify children with ASD by a multimodal behavior signal processing approach. The proposed system integrates autism domain knowledge and computer technologies, providing a standardized, objective, and interpretable tool for ASD diagnosis. The details of our system design are introduced in Sections III and IV.

III. Assessment Framework

A. Hardware Design

To minimize children's behavioral variance caused by outside factors, we design a testing studio to present standardized and objective audio-visual stimuli and capture the participants' behavioral data into RGB-D and audio signals.

Fig. 2(a) shows the appearance of the testing studio, a box-like soundproof cube with a side length of 3.2 and a height of 2.8 meters. Four high-definition short-focus projectors and two loudspeakers can display surround-screen videos with relatively stereo audio effects. During the assessment, children can get a relatively immersive experience of programmable audiovisual stimuli.

Fig. 2(b) demonstrates the deployment of multimodal sensors in the testing studio. We adopt a hybrid camera system to balance the imaging quality of RGB cameras and the equipment cost of depth sensors. First, 8 high-definition RGB cameras (LBAS-U3120-23C[1]) are used to capture videos at the resolution of 4096×3000. Second, 4 RGB-D cameras (Intel@Realsense-D455[2]) are installed at the room corners to provide RGB-D data at the resolution of 1280×720. All cameras are controlled to work simultaneously at 8 FPS. Also, wireless microphones (RODE Wireless GO II[3]) are equipped to record individuals' speech data. One is hidden in front of the child participant's seat,

[1]https://www.lusterinc.com/LBAS_Area_Scan_Camera_U3_/
[2]https://www.intelrealsense.com/depth-camera-d455/
[3]https://rode.com/cn/microphones/wireless/wirelessgoii#

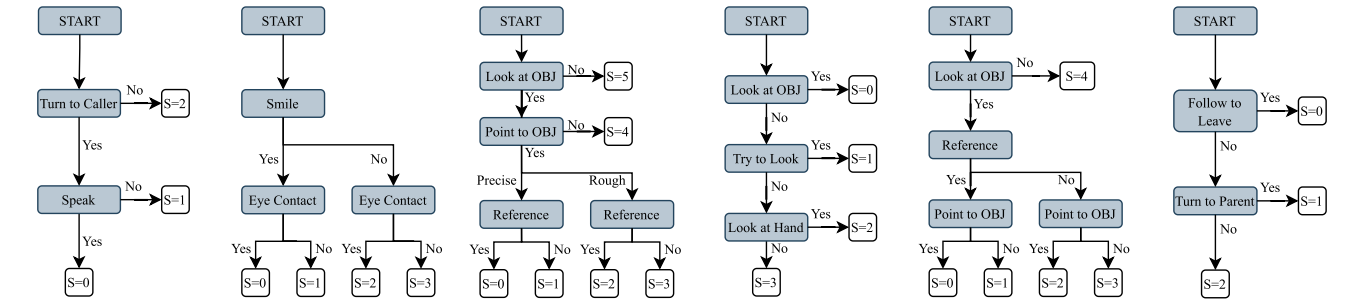


Fig. 3. Grading system for designed assessment paradigms. In general, the evaluation of each paradigm is based on the child participant's several fundamental behaviors: looking, pointing, responding, speaking, etc. The *Precise Point to OBJ* represents the child correctly pointing to the target object. The *Rough Point to OBJ* represents that the child performs the pointing gesture while he/she does not point in the correct direction (a large angle error). The *Reference* denotes the visual reference.

and the other two are with the assessor and the child's parent. Three synchronous audio streams are gathered at the sampling rate of 16 kHz.

We set up two desktop computers (Intel Core-i9 CPU, 16 GB Memory, and 1 TB SSD) to provide multimedia output and data acquisition, respectively. After each test, the recorded data will be uploaded to a computing workstation (Intel Core-i9 CPU, 256 GB Memory, 2 TB SSD, Two Nvidia 2080-TI GPUs) with Linux-Ubuntu 18.04 operating system. The subsequent behavior analysis on the workstation usually takes 3-4 hours for each test sample and then automatically output the analysis report as a PDF file.

In this way, the comprehensive multimodal behavioral data of the child, parent, and assessor can be gathered and analyzed user-friendly and cost-effectively. This testing studio can serve as a basic platform for generating programmable audio-visual stimuli and gathering behavioral data in a clinical setting.

B. Paradigm Design

To identify ASD children by analyzing their behavioral patterns, we present a set of assessment paradigms for quantifying children's social interaction skills from eye contact, pointing gestures, responding to joint attention, spoken language, and appropriate behavior [26]. In the testing studio, an assessor will lead the child and the parent to finish a series of interactive paradigms. Then, predefined grading rules will mark the child participant's assessment scores. The paradigms are introduced as follows.

1) Response to Name: Clinical research reveals that children who do not respond to their names by the age of 12 months have relatively high risks for autism, which can be a test with satisfactory specificity [40]. Here, we adopt the Response to Name (RN) paradigm to test children's reactions to being called names.

In the assessment, the child participant is first guided to play with toys on the desk. When his or her attention is drawn, the assessor suddenly calls the child's name from behind (position 225° depicted in Fig. 2(b)). If the child turns to face the caller with a language response, it will be marked as 0. If the child's response consists solely of looking without speaking, it will be marked as 1. Otherwise, no response will be marked as 2. Only in this case, the caller will repeat the name call after a 3-second pause.

Fig. 3(a) shows the grading rules for each name call. The max times of name calls are limited to 3, and the final score can be obtained by adding all scores in presented name calls. This paradigm has two sessions conducted by the assessor and the parent as the caller, respectively.

2) Social Smile: Studies of the Still-Face paradigm show that children at high risk of autism usually present fewer social smiles, which means the reduction of smiles in social activities can be a strong predictor of autism risk [41]. Therefore, we design the Social Smile (SS) paradigm to test children's ability to present smiles when receiving other's social stimuli.

This paradigm involves 4 sessions. The assessor and parent try to amuse the child by different methods. First, the assessor greets the child with a passional smile to test his or her feedback on others' positive emotions. Second, the assessor praises the child and evaluates his or her response to positive words. Third, the assessor plays an interactive game that tickles the child, which targets observing the child's reaction to slight body contact. The above sessions gradually increase the intensity of social stimuli, from distant greetings to verbal praise to physical touching. Moreover, the parent is asked to perform the last session by amusing the child in any way they usually do in their daily lives.

Fig. 3(b) shows the grading rules for each session. Based on whether the child participant can present a smiling response and deliver eye contact, this part of the assessment finally provides 4 individual scores ranging from 0 to 3.

3) Indicating Gesture: As a behavioral marker, the reduction of indicating gestures in social activities has been listed as an early warning symptom by the U.S. Centers for Disease Control and Prevention (CDC) [42]. Here, we introduce the Indicating Gesture (IG) paradigm to test children's ability to point to a target object in social interactions.

In the assessment, a Bluetooth controller triggers projectors to display a picture of a predefined target object on the walls of the testing studio. Then, the assessor begins to ask the child where the object is. In this way, we first examine the child's ability to understand verbal commands to look for the target object and then whether he or she will point to it and present a

visual reference (social reference). Specifically, visual reference is defined as the child sharing his or her findings with the assessor by eye contact after pointing to the object.

As the child participant may have a personalized preference for different objects, we select three cartoons that most kids may like: a red flower, a little tree, and a colorful balloon. The flower, tree, and balloon are placed at $315°$, $90°$, and $225°$ shown in Fig. 2(b). These settings lead to different levels of task difficulty. For instance, the child can easily find the flower if looking up a little. In contrast, he or she has to turn head back to find the balloon. Furthermore, we add a session to set the target object as the child's parent, targeting the performance of indicating gestures in social objects.

Fig. 3(c) shows the grading rules for each session. Based on whether the child participant can find the target object by looking at it, pointing to it, and initiating a visual reference, this assessment finally provides 4 individual scores ranging from 0 to 5.

4) Responding to Joint Attention: Children with ASD usually show few responses to joint attention in the early stage of development [43]. Many clinical diagnostic or screening tools (e.g., ESCS [44], ADOS [4]) have included this indicator. Thus, we design the Responding to Joint Attention (RJA) paradigm to test children's corresponding performance.

In the assessment, the assessor points to the clock mounted on the wall (position $180°$ in Fig. 2(b)) and asks the question, "Could you please tell me what time it is now?" Then, we examine whether the child can be attracted by the joint attention initiated by the assessor and understand the verbal commands to finish the tasks successfully.

Fig. 3(d) shows the grading rules for this paradigm. The *Try to Look* status represents the child looking around to seek the clock while not finding the correct direction. If the child keeps staring at the assessor's outstretched hand without seeking anything, the *Look at Hand* status will represent that the child participant can not understand the assessor's words and body language. This single-session paradigm finally provides one score ranging from 0 to 3.

5) Initiating Joint Attention: Observing children's ability to initiate joint attention with others is an important part of clinical assessment [44], often used in conjunction with the RJA. Here, we design the Initiating Joint Attention (IJA) paradigm to test children's ability to initiate joint attention in social interactions.

In the assessment, the child participant is guided to play with toys. The wall screen and the stereo audio device will suddenly display a three-second animation to attract the child's attention. Then, we examine the child participant's spontaneous reaction to the object that suddenly appears.

This paradigm contains 3 sessions with different audio-visual cartoon materials. In the first session, a yellow bird flapping its wings suddenly appears on the left wall (position $270°$ depicted in Fig. 2(b)). For the second session, a cartoon car with spinning wheels is presented to the child's right side (position $90°$ depicted in Fig. 2(b)). In the third, a cartoon cow waving its ears arises on the right rear wall (position $135°$ depicted in Fig. 2(b)). When each animation plays, the loudspeaker closest to the location of visual content will output the related sound material: birds chirp, car beeping, and cow moo.

TABLE I
SUMMARY OF DESIGNED ASSESSMENT PARADIGMS

Paradigm	Session	Initiator	Target Object	Score Range
RN	P1	Assessor	-	$[0, 6]$
	P2	Parent	-	$[0, 6]$
SS	P3	Assessor	-	$[0, 3]$
	P4	Assessor	-	$[0, 3]$
	P5	Assessor	-	$[0, 3]$
	P6	Parent	-	$[0, 3]$
IG	P7	Assessor	Flower	$[0, 5]$
	P8	Assessor	Tree	$[0, 5]$
	P9	Assessor	Balloon	$[0, 5]$
	P10	Assessor	Parent	$[0, 5]$
RJA	P11	Assessor	Clock	$[0, 3]$
IJA	P12	-	Bird	$[0, 4]$
	P13	-	Car	$[0, 4]$
	P14	-	Cow	$[0, 4]$
SA	P15	Parent	-	$[0, 2]$
	P16	Parent	-	$[0, 2]$

Fig. 3(e) shows the grading rules for each session. Based on whether the child participant can find the object and initiate joint attention by presenting the visual reference and pointing gesture, this part of the assessment finally provides 3 individual scores ranging from 0 to 4.

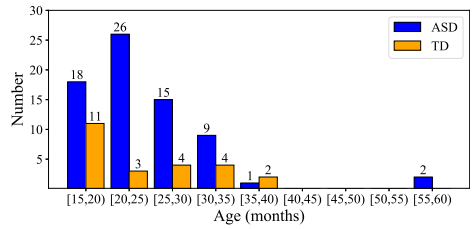
6) Separation Anxiety: Since toddlers usually have strong attachments to their parents, we utilize this phenomenon to design the Separation Anxiety (SA) paradigm for testing children's sensitivity to the absence of their parents.

As the last part of the assessment, this paradigm requires the parent to go out of the testing studio along the predefined leaving path shown in Fig. 2(b). This idea lets the parent walk in front of the child and explicitly makes the leaving visible to the child. After that, we observe the child participant's response to the parent's absence.

Fig. 3(f) shows the grading rules for this paradigm. If the child obtains a score of 2, we will introduce an additional session with more intensive language stimulation. The parent will call the child's name outside the door and say, "Hi, mom is leaving. You have to play alone." Then, the grading rules can be used again to score the additional session.

7) Summary of the Paradigm Design: The assessment proceeds smoothly and naturally. In the beginning, the assessor and parent call the child's name (RN) and greet the child with smiles (SS). After a warmup, high-level interactions (IG and RJA) begin. When the child is familiar with the social environment, we test whether he or she can initiate joint attention (IJA). Finally, all participants can leave the testing studio in the last paradigm (SA).

Table I summarizes the structured assessment for comprehensively quantifying children's social interaction skills. In a complete assessment, the child participant will receive 16 scores reflecting his or her performance in different paradigm sessions. The paradigms all follow the same standard: a lower score



(a) Age Distribution

(b) Gender Distribution

Fig. 4. Statistics of ASD/TD groups in the clinical database. Barplot graphs demonstrate the age and gender distributions of children with ASD and TD.

represents more typical behavior, and a higher score represents higher autism risk.

C. Database Collection

The collection and analysis of the clinical database are approved by the Institutional Review Board (IRB) of the Third Affiliated Hospital of Sun Yat-sen University and Duke Kunshan University. We recruit Chinese children from hospital clinics and social communities. A total of 116 children's parents agree to participate in this study. After excluding children who cannot complete the experiments due to physical disability or genetic defects, the final database obtains 95 participants between 16-56 months old, with an average of 24.87 months old. There are 71 children diagnosed with Autism Spectrum Disorder (ASD). The remaining 24 children are the normal group with Typical Development (TD). Fig. 4 illustrates the age and gender distributions of children with ASD and TD.

For each assessment case, the assessor will lead the child participant and parent to finish the designed assessment paradigms in the testing studio, usually taking 20 to 30 minutes. Then, our proposed computer system will automatically mark the child's paradigm scores which reflect the corresponding social interaction skills. The final computer-aided diagnostic suggestion will be given based on evaluated paradigm scores and behavior features.

After completing our computer-aided diagnosis session, each participant later took a formal clinical diagnosis process by several experienced clinicians with at least one chief physician. Furthermore, we invited the child participants to take the ADOS-2 [18] and ADI-R [14] assessments for ablation studies. Fig. 5 demonstrates the child participants' ADOS-2 and ADI-R score distributions in different scales, providing a fundamental portrait of children in our clinical database.

TABLE II
PERFORMANCE OF USED OPEN-SOURCE MODELS ON RELATED BENCHMARKS

Model	Dataset	Task	Metric (%)
SOLOv2	MS COCO [50]	Instance Segmentation	mAP=41.70
RetinaFace	FFDB [58]	Face Detection	mAP=99.22
ArcFace	LFW [59]	Face Recognition	ACC=99.52
BOTReID	Market1501 [60]	Person Re-Identification	mAP=95.00
HRNet	MS COCO [50]	Keypoint Detection	mAP=77.00

The abbreviations of mean average precision and accuracy are denoted as map and acc, respectively.

IV. Algorithm Framework

We propose a multimodal behavior signal processing framework to make the computer system mark children's paradigm scores automatically. During each assessment, the assessor will hold a Bluetooth controller to switch multimedia outputs and guide the paradigm progress. Therefore, the data acquisition system can record the paradigm timestamp to cut raw data into segments containing valid durations. Then, we design the behavior transcription system and response parser to convert multimodal signals into human behavioral data and calculate the child participant's responsive status in assessment paradigms. Finally, predefined scoring rules will mark the child's paradigm scores based on the recognized behavioral performance. The details of our framework design are introduced as follows.

A. Behavior Transcription

In the testing studio, the assessment process will be recorded into 8 high-resolution RGB videos, 4 low-resolution RGB videos with depth data, and 3-channel audio signals. We propose a Multiview and Multimodal Behavior Transcription (MMBT) system to recognize fundamental human behaviors from recorded data.

1) Multi-Person Identification and Localization: The core function of this module is to recognize participants' identities and localize their positional information in each frame of RGB videos. Several widely-used models are selected and organized into a pipeline to achieve the expected goal, shown in Fig. 6. Table II illustrates the performance of selected models on their respective benchmarks.

Each RGB video first goes through an instance segmentation model (SOLOv2 [45]) to extract human body regions (bounding boxes and masks). Meanwhile, the face detection model (RetinaFace [46]) and face recognition model (ArcFace [47]) extract face images to distinguish the identities of the child, parent, and assessor. The recognized body and face regions with the smallest intersection over union (IoU) will be matched. Furthermore, the body regions with known identities are collected across multiview videos to build each individual's body-region templates, which can be used for a person re-identification model (BOTReID [48]) to recognize identities of body regions without paired faces.

So far, the body and face regions extracted stay at pixel coordinates. The RGB-D videos with depth data will be further processed to localize the human skeleton, also known as body keypoints. Based on extracted body regions, we employ

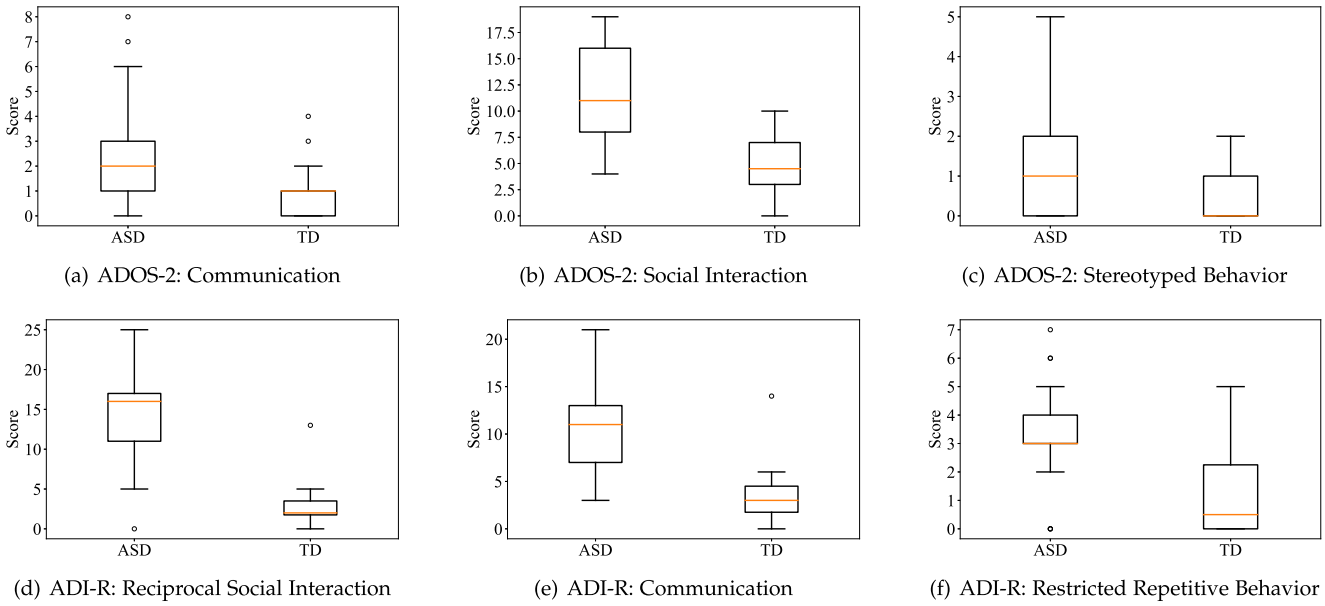


Fig. 5. Statistics of ASD/TD groups in the clinical database. Boxplot graphs of ADOS-2 scores depict the child participants' characteristics from the clinical point of view. The box stretches from the first quartile (Q1) to the third quartile (Q3) of each score distribution, with a colorful line representing the median. The whiskers extend the box to 1.5 times the interquartile range (IQR). Flier points denote scores beyond the end of the whiskers.

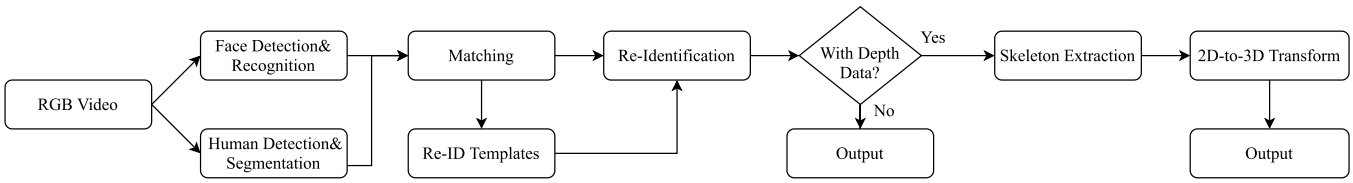


Fig. 6. Pipeline of multi-person identification and localization. After the processing, the identification and location data of individuals in each input video will be distinguished and extracted.

the HRNet model [49] to recognize 17 categories of human body keypoints defined in the MS COCO Dataset [50]. The position of a person's head will be calculated as the center of two eyes. Finally, using the corresponding depth data, the pixel coordinates of each person's 18 keypoints can be transformed into the physical three-dimensional coordinate system by the pinhole camera model [51].

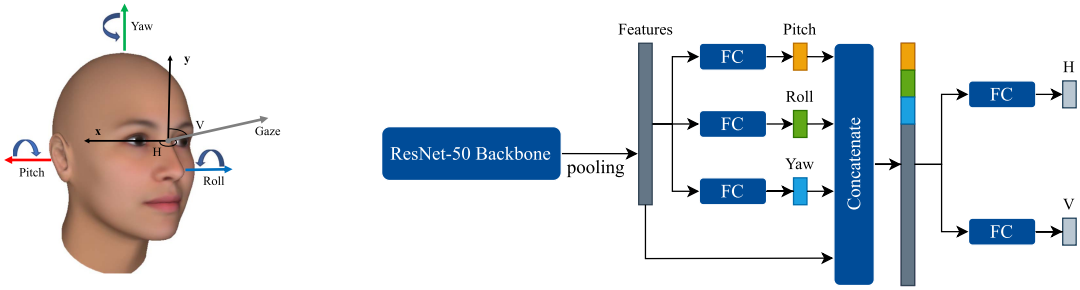
In our MMBT system, we introduce the term *Behavior Record* with the structure of an n-tuple to store all the recognized attributes of a specific person in a certain frame of a given video. The multi-person identification and location data in the RGB or RGB-D video can be abstracted into a sequence of behavior records. More attributes will be added to the behavior records in the following modules.

2) Gaze and Head Pose Estimation: Utilizing gaze data to analyze a participant's intentions can reveal rich information in clinical applications. However, common eye trackers (e.g., TOBII [52]) limit the participant to a very close distance and small range of head movement and rotation. Here, we propose a deep learning model to estimate the human gaze by RGB images. In the testing studio, high-definition cameras are arranged around the room to establish the camera-based gaze tracker system, which can support the surrounding gaze estimation in a contactless and unconstrained manner.

We adopt the SYSUGaze database [53] for training neural networks. The database contains 25,926 face images from 105 identities, with properties of large pose ranges, various lighting conditions, and diverse clothing styles. Moreover, it provides ground truth labels for both gaze and head pose, which are demonstrated in Fig. 7(a). The head pose is represented by pitch, yaw, and roll angles [54]. The horizontal and vertical angles describe the gaze direction.

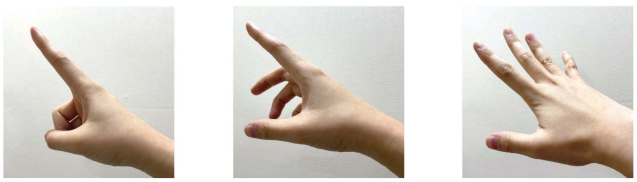
Fig. 7(b) shows the architecture of our proposed neural network model for jointly learning gaze and head pose estimation. The ResNet-50 model extracts a feature vector from the input face image. Then, three independent fully-connected layers (FCs) can predict the pitch, yaw, and roll angles of the head pose. The predicted head pose will be auxiliary information to concatenate with the original feature vector. The final two FCs are adopted to estimate vertical and horizontal gaze angles based on the newly constructed feature vector.

The SYSUGaze database is split into the training set (85 identities) and the test set (20 identities), with resizing each face image to 224×224. Based on the objective function used



(a) Representations of Gaze and Head Pose (b) Joint-Learning Model for Gaze and Head Pose Estimation

Fig. 7. Model architecture for gaze and head pose estimation. Based on the end-to-end deep neural networks, the angle-related data can be directly estimated from the face image.



(a) Gesture-1 (b) Gesture-2 (c) Gesture-3

Fig. 8. Predefined target gestures.

in [55], we train the proposed model by the Adam optimizer [56] with the learning rate of 1×10^{-4}. Experimental results show that our proposed method achieves the gaze estimation error of $4.87°$, a slight deviation between the estimated gaze direction and ground truth value. Moreover, it performs well in head pose estimation with a pitch error of $4.90°$, yaw error of $2.92°$, and roll error of $2.37°$.

The estimated head pose and gaze angles are finally transformed into the rotation matrix by Rodrigues' Rotation Formula [57]. With such a data-driven modeling approach, we utilize multiview cameras to build a surround eye tracker without constraints on participants in the testing studio.

3) Gesture Recognition: To extract each participant's hand regions in the recorded videos, we modify the output layer of the YOLOv5 [4] detection model to be a single-class hand detector. We train it on a large-scale mixed database with over 47,000 images, including OUHAND [61], Hand Dataset [62], Ego Hand [63], and TV/COCO-Hand [64]. The developed hand detector obtains the 64.86% mAP@95 on the test sets of adopted databases. Each detected hand region will be matched to the nearest wrist keypoint recognized in previous modules. Thus, the identities of extracted hand regions will not be confused.

We define three types of basic pointing gestures most common in clinics, shown in Fig. 8. The first one (Gesture-1) demonstrates the standard pointing gesture. The second one (Gesture-2) depicts an imperfect pointing gesture. In some cases, this situation may be caused by the limited muscular movement of children with developmental delays. The third one (Gesture-3) is a failure of the pointing gesture. Some children may be too young to make a correct pointing gesture but express their intentions by pointing their palms to an object. By adding a background category to represent the other cases, we design the visual gesture recognition in this work as a 4-class classification task.

TABLE III
STATISTICS OF THE MIXED GESTURE DATABASE. WE SELECT IMAGES FROM EACH SOURCE DATABASE AND RE-ORGANIZE THEM TO A LARGE-SCALE ONE

Source Database	Num. of Selected Images
American Sign Language [72]	8,700
Colombian Sign Language [73]	3,337
Indian Sign Language [74]	436
NUS Hand Posture - I [75]	240
NUS Hand Posture - II [76]	2,000
OUHands [61]	2,171
Jochen Triesch Static Hand Posture [77]	107
Total	16,991

As no existing database is suitable for our designed gesture taxonomy, we collect several related databases and re-assign their labels to meet our requirements, depicted in Table III. Also, we introduce an open gesture evaluation database [5] involving 2,850 hand images designed to test model performance outside the training data. All collected hand images are resized to 224×224 and fed to train a standard ResNet-50 model [65] with a few data augmentation methods (e.g., random rotation, colorization). The final gesture recognition system obtains an accuracy of 73.82% on the test set. Since the testing images usually have extremely complex backgrounds, the obtained accuracy can be regarded as satisfactory.

4) Facial Emotion Recognition: In this part, we construct a multi-task model for emotion recognition that can distinguish smile/non-smile facial expressions and estimate the valence and arousal values based on facial images. Therefore, the AffectNet database [66] is selected as training data, which involves approximately 450,000 images with two types of annotations: discrete facial expression classification (e.g., neutral, happy, sad) and continuous emotion regression of valence and arousal.

To further improve the database quality, we propose a processing pipeline to reduce samples with noisy annotations, shown in Fig. 9. First, we investigate three related databases with fewer images but high-quality annotations: ExpW [67] (91,793

[4]https://github.com/ultralytics/yolov5

[5]https://gitcode.net/EricLee/classification?from_codechina=yes

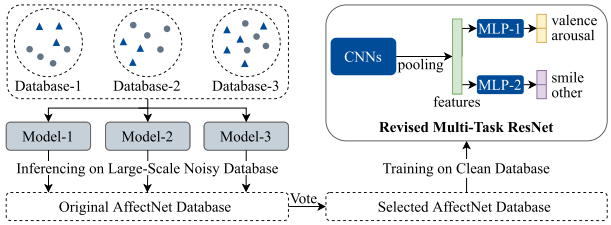


Fig. 9. Proposed data-centered processing pipeline for selecting high-quality samples from the original AffectNet dataset. The revised multi-task ResNet is trained on the selected database.

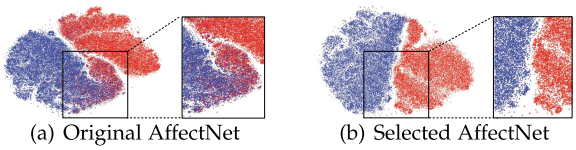


Fig. 10. The t-SNE visualization on different databases. Each point represents the embedding vector after t-SNE dimensionality reduction. The red and blue colors denote the smile and non-smile categories. Fig. 10(a) visualizes results by the model trained on the original AffectNet database. Fig. 10(b) shows the results of the model trained on the selected AffectNet database.

images), RAFDB [68] (29,672 images) and Genki-4K [69] (4,000 images). By setting all happy labels to the smile class and non-happy labels to the non-smile class, we build a mixed classification database (125,465 images) for the smile and non-smile facial expressions. Second, we split the mixed database into the train set and validation set at the ratio of $1:9$. Different model architectures (ResNet-50 [65], ResNeXt-50 [70], ResNet-18-ARM [71]) are trained to obtain accuracies of 99.4%, 92.67%, and 92.77% on the validation set, respectively. Based on ensemble learning, all images in the AffectNet database are individually classified by three models. Only samples with consistent predictions can be chosen to constitute a cleaned subset (233,419 images).

After the data cleaning, we obtain a new version of the AffecNet database assumed to have better annotation quality than the original one, which satisfies both large-scale data and multi-task annotation. We propose a multi-task model by modifying the ResNet-50 model [65] to have two task-oriented Multi-Layer Perceptions (MLPs) that can perform both classification and regression. By employing Alex Kendall's method [78] to weight loss functions with different units and scales in multi-task learning, the revised multi-task ResNet obtains an accuracy of 99.54% in the smile/non-smile classification task. Meanwhile, it has a mean squared error (MSE) of 0.2203 in valence regression and 0.2313 in arousal regression.

We implement the ablation experiment by training the proposed multi-task model on the original AffectNet database. The result shows a classification accuracy of 96.32%, MSE of 0.2771 in valence regression, and MSE of 0.2775 in arousal regression, proving the effectiveness of our data cleaning method. Moreover, the outputs after convolutional layers in the trained model are used to extract embedding vectors. Fig. 10 shows the t-SNE [79] visualization of the embedding vectors extracted by models trained on original and selected AffeNet databases. As

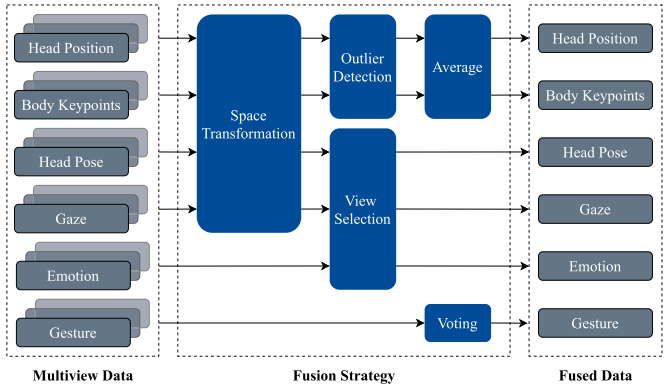


Fig. 11. Multiview fusion strategy. The multiview data of attributes recognized by vision-based models can be fused to achieve a more robust perception performance.

can be seen, the embedding representations from the selected database have a more precise classification boundary.

5) Multiview Fusion: The previous modules of the MMBT system have finished the vision-based perception to recognize human behaviors from each single-view camera. According to different task characteristics, we design a comprehensive strategy to fuse vision-perception results from the synchronized multiview data into one for better robustness, shown in Fig. 11.

For the spatial information (head position, body keypoints, head pose, and gaze), each point or vector is initially extracted in the respective camera coordinate system. Due to the varying camera placements, the first part of multiview fusion is to transform the above attributes from different coordinate systems to a unified one. We select the first RGB camera (position 0° depicted in Fig. 2(b)) as the world coordinate system of the testing studio. Then, a point in another coordinate system can be converted into the world coordinate system by a rigid transformation.

For example, let $\mathbf{p}_c \in \mathbb{R}^3$ represent the head position in a given camera coordinate system. It can be transformed into the world coordinate system by $\mathbf{p}_w = \mathbf{R} \times \mathbf{p}_c + \mathbf{T}$, where $\mathbf{R} \in \mathbb{R}^{3\times3}$ and $\mathbf{T} \in \mathbb{R}^3$ denote the rotation matrix and translation vector of the rigid transformation between two coordinate systems, which can be obtained by the chessboard calibration method [80]. This way, all spatial information recognized in different cameras is unified into the same world coordinate system.

The impulse noise caused by depth sensors often disturbs the localization of head position and body keypoints. Thus, we adopt the Isolation Forest algorithm [81] to detect outliers in candidates of each point transformed from different camera coordinate systems at the same timestamp. After discarding the outlier with the lowest confidence score in each frame, the remaining candidates will be averaged to obtain the fused results frame by frame.

As the recognition of head pose, gaze, and facial expression are entirely based on facial images with proper camera angles, we adopt the view-selection strategy for these attributes. At each timestamp, the best camera for each person will be chosen by recognizing the smallest yaw angle of the head pose. Each

person's head pose, gaze, and facial emotion come directly from the corresponding best camera.

The last part is a voting strategy to mitigate unreliable gesture recognition caused by occluded hands. We set a high decision threshold for the gesture recognition model to decrease the false-positive rate in a single camera. To guarantee the overall false-negative rate, once more than two of all surrounding cameras capture the pointing gestures simultaneously, the fused result is set to the target one.

6) Speech Recognition: Beyond the multiview RGB-D signals, we also implement an Automatic Speech Recognition (ASR) system to extract participants' speaking contents in the testing studio. The open-source toolkit Kaldi [82] is adopted due to its high usability and efficiency.

We collect several Mandarin corpora and re-train the Kaldi-based ASR system to recognize Chinese conversations. The first corpus is AISHELL-2 [83], which contains over 1,000 hours of indoor speech data from 1991 speakers, recorded at a sample rate of 16 kHz. Additionally, we include several large-scale open corpora hosted at openslr.org with approximately 6,953 hours of speech data speaking by 5,590 identities. Based on the mixed speech corpora, our Kaldi-based ASR system achieves a character error rate (CER) of 1.37% on AISHELL-1 [84] and 3.2% on AISHELL-2 [83] benchmarks, respectively.

B. Response Parser

Based on the attributes extracted from the MMBT system, we build a response parser module to determine whether a predefined response happens during the assessment. According to the paradigm design, we define some basic response categories that should be detected.

1) Look at Object: To determine whether the child is looking at the target object at timestamp t, we obtain the child's gaze vector $\mathbf{g}_t \in \mathbb{R}^3$, the head position $\mathbf{p}_t \in \mathbb{R}^3$, and the target object position $\mathbf{o}_t \in \mathbb{R}^3$ from the behavior transcription. We apply head pose instead of gaze data when the gaze estimation is unavailable due to occlusion. Regarding the head position (center of two eyes) as the starting point of the gaze ray, the angle θ between gaze direction and object direction can be calculated by

$$\theta_t = \arccos \frac{\mathbf{g}_t \cdot (\mathbf{o}_t - \mathbf{p}_t)}{\|\mathbf{g}_t\|_2 \times \|\mathbf{o}_t - \mathbf{p}_t\|_2}. \tag{1}$$

Let E_1 represent the event that the child is looking at the target object, we define the probability of E_1 at timestamp t as follows:

$$P_t(E_1) = \frac{\pi - \theta_t}{\pi}. \tag{2}$$

Once the probability exceeds a preset threshold, the event of "Looking at Object" can be detected to happen.

The target object can be used in different aspects. When setting it to the assessor or parent, E_1 represents that the child responds to the social interaction by turning his or her head to look at the other person. If the child and the assessor/parent look at each other synchronously, this situation can be considered eye contact. When setting the target object to cartoons (e.g., red flower, little tree, colorful balloon, bird, car) displayed on walls, it is regarded that the child looks in the direction of placing these objects.

2) Point to Object: There are two steps to decide whether a hand is pointing to the target object. First, the child's hand must present one of the pointing gestures (Gesture-1 and Gesture-2 shown in Fig. 8). Second, the angle between the pointing direction and object direction must be small enough.

We define the extension line from the child's elbow to the wrist as the pointing direction of the hand. Let $\mathbf{e}_t \in \mathbb{R}^3$, $\mathbf{w}_t \in \mathbb{R}^3$, and $\mathbf{o}_t \in \mathbb{R}^3$ denote the elbow, wrist, and target object coordinates at timestamp t. The angle α between the pointing direction and the object direction at timestamp t can be calculated by

$$\alpha_t = \arccos \frac{(\mathbf{o}_t - \mathbf{e}_t) \cdot (\mathbf{w}_t - \mathbf{e}_t)}{\|\mathbf{o}_t - \mathbf{e}_t\|_2 \times \|\mathbf{w}_t - \mathbf{e}_t\|_2}. \tag{3}$$

Let E_2 represent the event that the child participant's hand is pointing to the target object. We define the probability of E_2 at timestamp t as follows:

$$P_t(E_2) = \frac{\pi - \alpha_t}{\pi} \times P_t(G), \tag{4}$$

where $P_t(G) \in [0, 1]$ indicates the probability that the corresponding hand is recognized as the pointing gesture. This policy ensures that E_2 can be determined to happen only if the child points in the correct direction and expresses a pointing gesture at the same time.

We calculate the $P_t(E_2)$ for the child's two hands over all timestamps. Once there is a moment that one of the two pointing probabilities exceeds a preset threshold, the event of "Point to Object" can be detected to happen. If the child only shows a pointing gesture without pointing in the correct direction, this situation will be considered a "Rough Point to Object" event introduced in Fig. 3(c)).

3) Smile: To determine whether the child is smiling at timestamp t, we obtain the child's facial expression category (smile/non-smile), valence, and arousal from the MMBT system. We adopt the estimated emotional valence to adjust the decision sensitivity rather than only using the categorical label to provide a binary decision.

Let E_3 denote the event that the child participant is smiling. We define the probability of E_3 at timestamp t as follows:

$$P_t(E_3) = P_t(S) \times P_t(V), \tag{5}$$

where $P_t(S) \in [0, 1]$ and $P_t(V) \in [0, 1]$ are the confidence score of the smile category and emotional valence estimated by the MMBT system. Once the probability exceeds a preset threshold, the event of "Smile" can be detected to happen. As valence refers to the pleasantness of an emotional state, the computation of $P_t(E_3)$ incorporates both the occurrence of a smile expression and its emotional intensity.

4) Speak: Based on automatic speech recognition (ASR), the computer system can search for target keywords (e.g., child's name, assessor's questions) occurring in each participant's microphone data. We modify the ASR outputs to build a keyword spotting (KWS) system for detecting the occurrences of a given keyword.

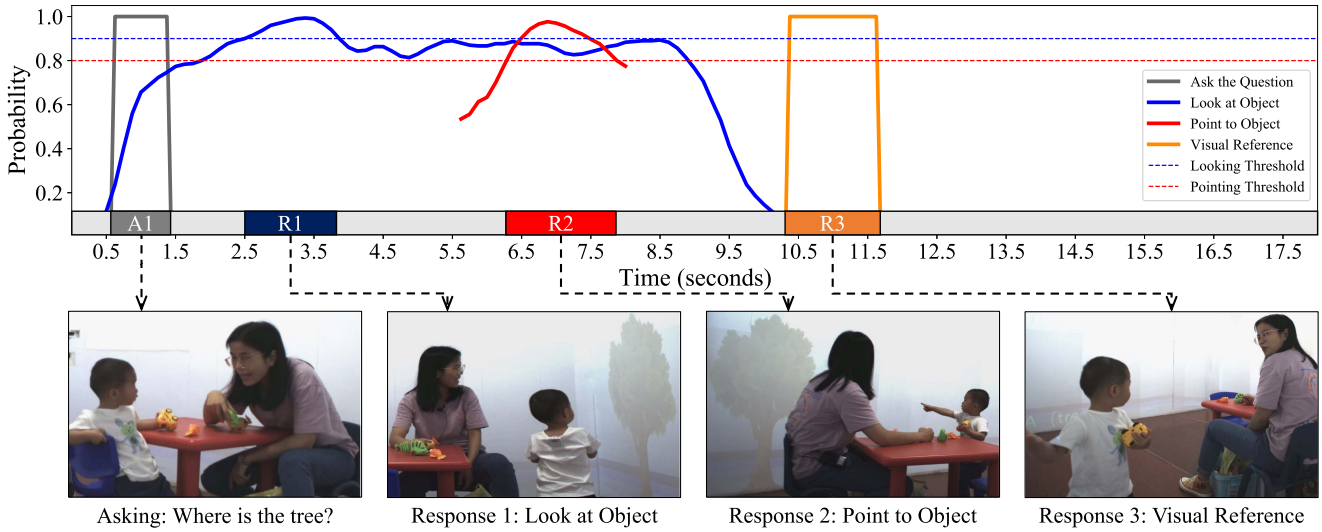


Fig. 12. Demonstration of rule-based paradigm scoring. The line chart represents the probabilities of different responses (events) occurring during the assessment. In the beginning, the assessor asks the child: "Where is the tree? Could you please help me find it?" One second later, the child looks at the tree displayed on the wall. Then, he presents a pointing gesture in the direction of the tree. After a brief pause, the child turns his head back to share findings with the assessor, which is recognized as a visual reference. According to the predefined grading rules, the score of this case can be marked as 0.

However, it is hard to match the child's name from the original ASR outputs due to Chinese homophones. Hence, we first convert the ASR predictions from word to phoneme by MDBG Chinese Dictionary[6]. Then, Levenshtein distance [85] with a shifted window is used to measure the difference between phonemes of window content and the keyword. The keyword occurrence can be detected once the Levenshtein distance is lower than a threshold.

With this KWS system, our system can acquire the name-calling timestamps in the Response to Name paradigm. Similarly, in the Indicating Gesture paradigm, the start time when the assessor begins to ask questions can be known. If only considering the presence of spoken words in the child's audio stream, it can be used to determine whether the child presents a verbal response.

5) Leave: The Separation Anxiety paradigm relies on detecting whether the child follows his or her parent to leave the testing studio. Let E_4 represent the event that a target person walks out of the testing studio. The closer the person is to the door, the more likely he or she is to leave. As coordinates of each person's positions can be obtained from the MMBT system, we define the probability of event E_4 at timestamp t as follows:

$$P_t\left(E_4\right) = 1 - \frac{\min(\left\|\mathbf{p}_t - \mathbf{o}_d\right\|_2,\ D)}{D}, \tag{6}$$

where $\mathbf{p}_t$ denotes the target person's coordinates; $\mathbf{o}_d$ denotes the door position in the testing studio. D represents the maximum diagonal length of the testing studio to normalize $P_t(E_4) \in [0, 1]$. Once the probability exceeds a preset threshold, the computer system can obtain the timestamps of the parent or child leaving the testing studio.

6) Rule-Based Scoring: By the response parser, participants' behaviors in the testing studio can be summarized into several basic categories: *Look at OBJ*, *Point to OBJ*, *Smile*, *Speak*, and *Leave*. Each item of them needs several preset thresholds to determine the occurrence of a target behavior. As the testing studio has a wide range of activity space, those thresholds do not rely on overly strict settings. Before the formal database collection, we test a few preliminary samples and set the thresholds empirically by observing these case studies. Finally, the assessment score of each paradigm can be easily marked by corresponding tree-like grading rules. Fig. 12 shows an example of scoring the second session of the Indicating Gesture paradigm.

V. Experimental Results

Our designed diagnosis pipeline starts by leading children to finish a series of assessment paradigms. Then, their social interaction skills can be estimated based on their performance during the assessment, and the final diagnostic suggestion will be further determined. Therefore, we evaluate the proposed computer-aided ASD diagnosis system from three aspects: the performance of paradigm scoring, the effectiveness of paradigm design, and the use of machine learning in ASD diagnosis.

A) Evaluation of Paradigm Scoring

Based on the proposed multimodal behavior signal processing framework, each child participant can obtain 16 assessment scores to quantify his or her social interaction skills in multiple aspects. $\mathbf{S_{comp}}$ denotes the collection of all children's paradigm scores in the clinical database, which is given by our computer system automatically.

Moreover, two professionals with over three years of clinical experience review the recorded videos to mark the paradigm scores independently, providing the scoring results $\mathbf{S_{anno1}}$ and $\mathbf{S_{anno2}}$. Then, inconsistent labels will be retrospected to address the controversial annotations, resulting in the revised scoring

[6]https://www.mdbg.net/chinese/dictionary

TABLE IV
PERFORMANCE OF PARADIGM SCORING BY THE COMPUTER SYSTEM AND TWO ANNOTATORS

Paradigm	Level	Range	$\mathbf{S_{comp}}$			$\mathbf{S_{anno1}}$			$\mathbf{S_{anno2}}$		
			ACC (%)	MAE	Kappa	ACC (%)	MAE	Kappa	ACC (%)	MAE	Kappa
Response to Name (RN)	7	$[0, 6]$	73.68	0.6105	0.6222	94.18	0.1164	0.9148	92.59	0.1852	0.8913
Social Smile (SS)	4	$[0, 3]$	69.17	0.4424	0.5508	75.07	0.3686	0.6355	83.65	0.2440	0.7576
Indicating Gesture (IG)	6	$[0, 5]$	66.49	0.6383	0.4621	83.33	0.2715	0.7514	80.32	0.3191	0.6899
Responding to Joint Attention (RJA)	4	$[0, 3]$	72.34	0.4574	0.6179	79.35	0.2391	0.7211	87.10	0.1398	0.8212
Initiating Joint Attention (IJA)	5	$[0, 4]$	76.68	0.3852	0.6894	82.08	0.2688	0.7625	80.21	0.3463	0.7346
Separation Anxiety (SA)	3	$[0, 2]$	89.87	0.1203	0.8449	94.19	0.0814	0.9090	91.82	0.1006	0.8703

The statistics are calculated by setting the svote as the ground truth. acc, mae, and kappa are the abbreviations of accuracy, mean absolute error, and cohen's kappa coefficient.

TABLE V
COMPARISON OF PARADIGM SCORING BETWEEN TWO ANNOTATORS

Paradigm	Level	Range	ACC (%)	MAE	Kappa
RN	7	$[0, 6]$	88.83	0.2287	0.8365
SS	4	$[0, 3]$	64.77	0.5068	0.4829
IG	6	$[0, 5]$	69.62	0.4866	0.5338
RJA	4	$[0, 3]$	70.33	0.3077	0.5989
IJA	5	$[0, 4]$	70.61	0.4444	0.6073
SA	3	$[0, 2]$	86.54	0.1667	0.7865

The statistics are calculated by setting one of the annotators as ground truth and the other as testing data.

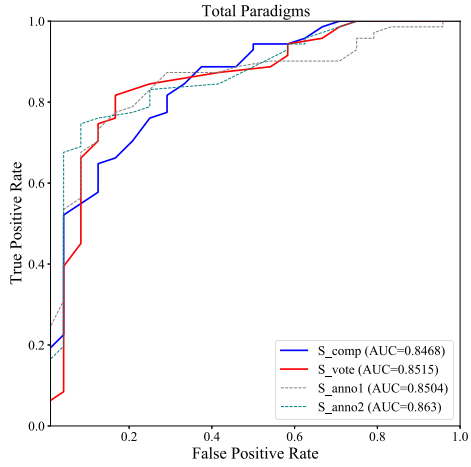


Fig. 13. Receiver operating characteristic (ROC) curve of ASD classification based on the sum of 16 scores in all paradigms.

results $\mathbf{S_{vote}}$. Table IV evaluates the scoring results of the computer system and two annotators from both classification and regression perspectives.

The highest accuracy for the computer system is 89.87% on the Separation Anxiety (SA) paradigm, which is close to the two annotators (94.19% and 91.82%). The Response to Name (RN) paradigm is easy for two annotators to obtain 94.18% and 92.59% accuracy. However, the computer cannot localize name-calling timestamps in occasional dialects as accurately as humans. Our proposed computer system only has an accuracy of 73.68%. For the Responding to Joint Attention (RJA) and Initiating Joint Attention (IJA) paradigms, the computer system shows acceptable accuracies not much worse than human performance.

The computer system has relatively poor accuracy in scoring the paradigms of the Social Smile (69.17%) and Indicating Gesture (66.49%). Besides the complex scoring levels, subjectivity is a common problem in both tasks. For instance, the decision-making for the smile and non-smile appearance is highly confusing in the SS paradigm. Furthermore, the IG paradigm needs to distinguish between precise and rough pointing, which is more ambiguous than only finding the pointing gesture occurrence.

Mean absolute error (MAE) can also provide an evaluation from the regression perspective. The lower the MAE value, the better performance it represents. The computer system obtains its top-2 worst results in the IG and RN paradigms: 0.6383 and 0.6105. Although the IG and RN paradigms have a relatively larger scoring range than others, their MAE values are still smaller than 1. Moreover, the computer system achieves an MAE value of less than 0.5 in all other items, which indicates that the misclassified computer scores are close to the actual values.

Table V compares the inconsistency between the two annotators. We introduce Cohen's kappa coefficient for measuring the inter-rater agreement for categorical variables. Apart from the RN and SA paradigms, the kappa value between $\mathbf{S_{comp}}$ and $\mathbf{S_{vote}}$ in Table IV does not show significant disadvantages compared to human-to-human consistency.

B) Evaluation of Paradigm Effectiveness

We adopt the Receiver Operating Characteristic (ROC) curve to test whether the assessment paradigms contribute to the ASD diagnosis. Specifically, once the total score of given paradigms exceeds a predefined threshold, the child participant is classified into the ASD group. By traversing different thresholds, we analyze the ROC curves and the areas under the ROC curves (AUC) of individual paradigms and their combination.

Fig. 13 shows the ROC curve based on the sum of 16 scores in all paradigms. The computer system obtains an AUC of 0.8468, comparable to the human-scoring results. Moreover, the $\mathbf{S_{vote}}$ does not obtain the highest AUC value, which means the most objective scoring does not necessarily lead to the best ASD identification. Although the paradigm scoring between the computer system and two annotators has an accuracy gap, they still show relative performance in the ROC analysis.

Fig. 14 shows the AUC values of individual paradigms. The human-labeled $\mathbf{S_{vote}}$, $\mathbf{S_{anno1}}$, and $\mathbf{S_{anno2}}$ have identical

TABLE VI
PEARSON CORRELATION COEFFICIENTS (PCC) BETWEEN OUR DESIGNED PARADIGM, ADOS-2, AND ADI-R SCORES

Paradigm Score	ADOS-2				ADI-R			
	Commu-nication	Social Interaction	Stereotyped Behavior	Total	Reciprocal Social Interaction	Commu-nication	Restricted Repetitive Behavior	Total
$\mathbf{S_{comp}}$	0.5133	0.6790	0.3889	0.6856 ($P < 0.001$)	0.6865	0.5234	0.3271	0.6383 ($P < 0.001$)
$\mathbf{S_{vote}}$	0.5524	0.6999	0.4720	0.7337 ($P < 0.001$)	0.6872	0.4690	0.2846	0.6100 ($P < 0.001$)
$\mathbf{S_{anno1}}$	0.5271	0.7223	0.4690	0.7568 ($P < 0.001$)	0.6803	0.4527	0.2764	0.5983 ($P < 0.001$)
$\mathbf{S_{anno2}}$	0.5228	0.7042	0.4859	0.7446 ($P < 0.001$)	0.6447	0.4547	0.2694	0.5788 ($P < 0.001$)

The ados-2 and adi-r have three scales in different assessment aspects. the paradigm scoring can be obtained by the computer system, the two annotators, and their votes. P denotes the p-values for testing non-correlation under a 0.95 confidence interval (CI).

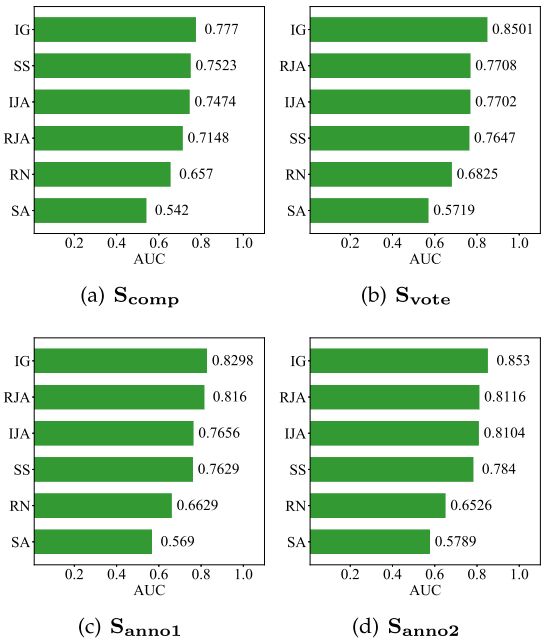


Fig. 14. Area under the ROC curve (AUC) rankings of individual paradigms obtained by different versions of paradigm scoring.

rankings. Differently, the Social Smile (SS) paradigm scored by the computer system has a relatively higher AUC ranking. We speculate that this is due to the greater consistency of the computer system in recognizing facial expressions, as opposed to human subjectiveness.

Furthermore, we compare our paradigm design with existing clinical toolkits: ADOS-2 [18] and ADI-R [14]. Table VI illustrates the Pearson Correlation Coefficients (PCC) between the child participants' total score of our paradigms and the multiple scales of ADOS-2 and ADI-R assessments. Regardless of the scoring methods, the child participants' paradigm scores show clear positive correlations with the scales of Social Interaction in ADOS-2 and Reciprocal Social Interaction in ADI-R. As our paradigm design targets the evaluation of social interaction skills, the correlations between our paradigm scores and the other scales in ADOS-2 and ADI-R are not significant enough.

C. Evaluation of ML-Based ASD Diagnosis

1) Diagnosis by Paradigm Scores: After completing the assessment, a child participant will obtain 16 scores in all paradigm sessions. Based on the 16-dimensional feature vector built on paradigm scores, three widely-used machine learning classifiers with small-size data are adopted as benchmark models: Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB). Children with ASD are regarded as positive samples for the binary classification task. On the contrary, TD children are grouped into a negative category. We train and evaluate the models by the leave-one-out cross-validation method.

Table VII shows the performance of adopted models trained on different versions of paradigm scorers. None of the three classifiers express a dominant advantage over the others, which shows that the paradigms are well-designed to integrate with ASD clinical experience. The size of the clinical database is not large enough to bring significant discrepancies for those data-driven models.

Comparing different versions of paradigm scores, the $\mathbf{S_{comp}}$ leads to the SVM model reaching the highest accuracy of 84.21% and the F1-score of 89.21%. NB model has a stable performance to obtain an accuracy of over 80% in all kinds of input data. $\mathbf{S_{vote}}$ shows relatively poor performance than $\mathbf{S_{anno1}}$ and $\mathbf{S_{anno2}}$. Again, the most objective scoring results do not have the best effects on the ASD diagnosis. Since two annotators are ASD specialists with years of professional experience, their scoring processes will inevitably involve more subjective judgments according to the participant's overall performance (e.g., vocal mood, micro facial expression and body language). Therefore, the specialist-scored results may imply more rich information than rigid computer-scored rules.

2) Diagnosis by Quantitative Behavior Features: We further improve the computer system by introducing quantitative behavior features that are not available for human scoring. Based on the recognized human behaviors with timestamps, we define three kinds of indicators:

- *Latency:* the time gap between the start of a social stimulus and the child's response. For example, if the parent calls the child's name at time t_0 and the child looks at the caller at time t_1, the latency of this response will be $t_1 - t_0$.
- *Duration:* the period of a specific response. For example, if the child's smile begins at time t_1 and ends at time t_2, the duration of this response will be $t_2 - t_1$.
- *Intensity:* the proportion of a response duration in a paradigm session. In the Social Smile paradigm, it is difficult to determine when the assessor begins to amuse the child participant. Hence, intensity is adopted instead of latency.

TABLE VII
PERFORMANCE OF LOGISTIC REGRESSION (LR), SUPPORT VECTOR MACHINE (SVM), AND NAIVE BAYES (NB) MODELS TRAINED ON DIFFERENT VERSIONS OF PARADIGM SCORING

Input Data	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Sensitivity (%)	Specificity (%)
$\mathbf{S_{comp}}$	LR	80.00	84.21	90.14	87.07	90.14	50.00
	SVM	84.21	91.18	87.32	89.21	87.32	75.00
	NB	80.00	88.24	84.51	86.33	84.51	66.67
$\mathbf{S_{vote}}$	LR	77.89	82.89	88.73	85.71	88.73	45.83
	SVM	77.89	86.76	83.10	84.89	83.10	62.50
	NB	80.00	90.62	81.69	85.93	81.69	75.00
$\mathbf{S_{anno1}}$	LR	78.95	83.12	90.14	86.49	90.14	45.83
	SVM	80.00	85.14	88.73	86.90	88.73	54.17
	NB	82.11	90.91	84.51	87.59	84.51	75.00
$\mathbf{S_{anno2}}$	LR	77.89	83.78	87.32	85.52	87.32	50.00
	SVM	82.11	88.57	87.32	87.94	87.32	66.67
	NB	81.05	90.77	83.10	86.76	83.10	75.00

TABLE VIII
SUMMARY OF ADOPTED QUANTITATIVE INDICATORS FOR EACH ASSESSMENT PARADIGM

Paradigm	Response	Indicator		
		Latency	Duration	Intensity
RN	Turn to Caller	✓	✓	
SS	Smile		✓	✓
	Eye Contact		✓	✓
IG	Look at OBJ	✓	✓	
	Point to OBJ	✓	✓	
	Visual Reference	✓	✓	
RJA	Look at Hand	✓	✓	
	Look at Clock	✓	✓	
IJA	Look at OBJ	✓	✓	
	Point to OBJ	✓	✓	
	Visual Reference	✓	✓	
SA	Turn to Parent	✓	✓	
	Follow to Leave	✓		

Each response can be described by one or two quantitative indicators according to different paradigm characteristics.

Table VIII illustrates how we represent the original responses by different quantitative indicators. Each essential response can be described as a subset of latency, duration, and intensity. As the age distribution of children in our clinical database is relatively young, there are too many missing values for the *Speak* response. Here we drop this feature. Moreover, the parent's leaving means the end of the assessment process, and there is no duration for the subsequent behaviors. The *Leave* response is just described by latency. In this way, each child participant's assessment performance can be described by a 72-dimensional feature vector representing the latency, duration, and intensity of responses in paradigm sessions.

The collection of all children's quantitative features is denoted as $\mathbf{F_{comp}}$. Each feature dimension is normalized to $[0, 1]$. Sometimes, the child participant may not respond to external stimuli at all. For these situations where quantitative features cannot be calculated, we directly set the default latency to 1 while duration and intensity to 0, reflecting the worst social response. Furthermore, we incorporate the paradigm scores and quantitative features to build a new set $\mathbf{S_{comp}} \cup \mathbf{F_{comp}}$, which aims to take advantage of both.

Table IX evaluates the models trained on $\mathbf{F_{comp}}$ and $\mathbf{S_{comp}} \cup \mathbf{F_{comp}}$. In most evaluation metrics, quantitative features outperform score-based features by introducing continuous variables that may contain more fine-grained information than discrete paradigm scores. In addition, the system performance can be further improved by combining quantitative and score-based features, with confusion matrices depicted in Fig. 15. The SVM model trained on $\mathbf{S_{comp}} \cup \mathbf{F_{comp}}$ achieves the highest accuracy of 88.42% and F1-score of 92.20%. Although the collected database has imbalanced ASD/TD groups, the best model still obtains a sensitivity of 91.55% and specificity of 79.17%, representing a competitive performance.

3) Discussions: Experimental results of Tables VI and VII demonstrate the validity of our paradigm design in ASD diagnosis, which proves the links between ASD risks and atypical behavioral patterns (e.g., visual attention [22], [23], orienting to name call, social reference, and responsive social smile [24], [25], [26], [27]). In addition, Table IX shows that advanced improvements can be obtained by quantitative behavioral features extracted by the computer system. Our proposed Behavior Signal Processing (BSP) framework is a more comprehensive and practical solution than previous methods [6], [9], [28]. Unfortunately, the current BSP framework still lacks the evaluation of participants' stereotypical motor movement. We will study how to add this function unconstrainedly in the version.

To analyze the diagnosis results of the obtained best model, we inspect the medical history of the 11 participants misclassified in the SVM confusion matrix shown in Fig. 15. (1) The false-negative participants have an average age of 26.36 months old. One participant has accepted intervention for half a year before joining our experiments. Another one is the high-risk cohort whose sibling has been diagnosed with ASD. Both had mild symptoms, which are complex samples for clinical diagnosis. (2) The false-positive participants have an average of 19.9 months old. All of them are no older than 24 months old. In contrast, all the false-negative participants are not younger than 24 months old. It shows that the behaviors of too young children are often affected due to incomplete body development. For

TABLE IX
COMPARISON OF LOGISTIC REGRESSION (LR), SUPPORT VECTOR MACHINE (SVM), AND NAIVE BAYES (NB) MODELS TRAINED ON QUANTITATIVE FEATURES AND THE COMBINATION WITH PARADIGM SCORES

Input Data	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Sensitivity (%)	Specificity (%)
$\mathbf{F_{comp}}$	LR	83.16 (+3.16)	87.67 (+3.46)	90.14	88.89 (+1.82)	90.14	62.50 (+12.5)
	SVM	86.32 (+2.11)	90.28 (-0.90)	**91.55** (+4.23)	90.91 (+1.70)	**91.55** (+4.23)	70.83 (-4.17)
	NB	82.11 (+2.11)	89.71 (+1.47)	85.92 (+1.41)	87.77 (+1.44)	85.92 (+1.41)	70.83 (+4.16)
$\mathbf{S_{comp}} \cup \mathbf{F_{comp}}$	LR	84.21 (+4.21)	87.84 (+3.63)	**91.55** (+1.41)	89.66 (+2.59)	**91.55** (+1.41)	62.50 (+12.5)
	SVM	**88.42** (+4.21)	**92.86** (+1.68)	**91.55** (+4.23)	**92.20** (+2.99)	**91.55** (+4.23)	**79.17** (+4.17)
	NB	81.05 (+1.05)	89.55 (+1.31)	84.51	86.96 (+0.63)	84.51	70.83 (+4.16)

In the brackets are the gaps to the baseline performance of models trained on paradigm scores only. the bold represents the best performance in each column.

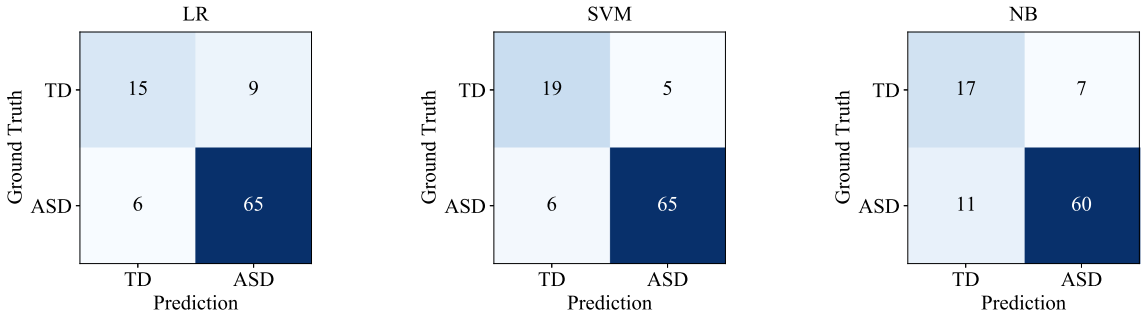


Fig. 15. Confusion matrices of Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB) models trained on the combination of paradigm scores and quantitative features from the computer system.

older children, the judgment of their behaviors should have some adjustments in the future.

We retrospect recorded videos of the misclassified participants. In some cases, the children can express appropriate responses, proving that they cannot be explained as a lack of social skills. However, the grading scope of our paradigm design fails to include all possible behaviors. These case studies show that it is difficult to deal with all ASD groups with the same standards. In future work, researchers would be inspired to consider multi-scale evaluation metrics for different groups (e.g., age, gender, medical history) and a broader scope of behavior analysis in the paradigm design.

VI. CONCLUSION

This paper proposes a computer-aided ASD diagnosis system integrating the hardware, algorithm, and ASD domain knowledge. In our clinical database of 95 participants with an average age of two, the computer system obtains its highest accuracy of 88.42% for ASD diagnosis, representing a comparable performance with human experts. The developed system has been deployed for routine tests in our partner hospital.

The most significant contribution of our work is to provide a standardized, objective, and programmable platform for stimulating, gathering, analyzing, modeling, and interpreting behavioral data in the application of computer-aided ASD diagnosis. Our work provides a replicable solution: behavioral data collected from different institutions can be analyzed jointly in a unified framework. In the future, the equipment will be duplicated to acquire valuable data continuously. Furthermore, we will further investigate end-to-end approaches for ASD diagnosis in the next stage.

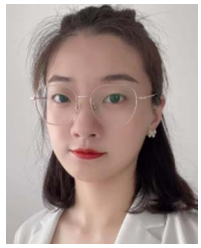
REFERENCES

[1] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders," 5th ed., 2013. [Online]. Available: https://doi.org/10.1176/appi.books.9780890425596
[2] M. J. Maenner et al., "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2018," *MMWR Surveill. Summaries*, vol. 70, no. 11, 2021, Art. no. 1.
[3] C. Lord et al., "The Lancet commission on the future of care and clinical research in autism," *Lancet*, vol. 399, no. 10321, pp. 271–334, 2022.
[4] P. C. DiLavore, C. Lord, and M. Rutter, "The pre-linguistic autism diagnostic observation schedule," *J. Autism Devlop. Disord.*, vol. 25, no. 4, pp. 355–379, 1995.
[5] T. Falkmer, K. Anderson, M. Falkmer, and C. Horlin, "Diagnostic procedures in autism spectrum disorders: A systematic literature review," *Eur. Child Adolesc. Psychiatry*, vol. 22, no. 6, pp. 329–340, 2013.
[6] J. Hashemi et al., "Computer vision analysis for quantification of autism risk behaviors," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 215–226, First Quarter 2021.
[7] Z. Zhao et al., "Identifying autism with head movement features by implementing machine learning algorithms," *J. Autism Devlop. Disord.*, vol. 52, pp. 3038–3049, 2022.
[8] Z. Zhao, H. Tang, X. Zhang, X. Qu, X. Hu, and J. Lu, "Classification of children with autism and typical development using eye-tracking data from face-to-face conversations: Machine learning model development and performance evaluation," *J. Med. Internet Res.*, vol. 23, no. 8, Aug. 2021, Art. no. e29328.
[9] T. Guha, Z. Yang, R. B. Grossman, and S. S. Narayanan, "A computational study of expressive facial dynamics in children with autism," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 14–20, First Quarter 2018.
[10] H. Y. Lee et al., "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 215–219.

[11] M. Li et al., "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Comput. Speech Lang.*, vol. 56, pp. 80–94, 2019.
[12] L. Sadouk, T. Gadi, and E. H. Essoufi, "A novel deep learning approach for recognizing stereotypical motor movements within and across subjects on the autism spectrum disorder," *Comput. Intell. Neurosci.*, vol. 2018, 2018, Art. no. 7186762.
[13] N. M. Rad et al., "Deep learning for automatic stereotypical motor movement detection using wearable sensors in autism spectrum disorders," *Signal Process.*, vol. 144, pp. 180–191, 2018.
[14] C. Lord, M. Rutter, and A. L. Couteur, "Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *J. Autism Devlop. Disord.*, vol. 24, no. 5, pp. 659–685, 1994.
[15] M. Dereu, *Modified Checklist for Autism in Toddlers (M-CHAT)*. New York, NY,USA: Springer, 2013, pp. 1890–1894. [Online]. Available: https://doi.org/10.1007/978-1-4419-1698-3277
[16] D. L. Robins, K. Casagrande, M. Barton, C.-M. A. Chen, T. Dumont-Mathieu, and D. Fein, "Validation of the modified checklist for Autism in toddlers, revised with follow-up (M-CHAT-R/F)," *Pediatrics*, vol. 133, no. 1, pp. 37–45, 2014.
[17] E. Rellini, D. Tortolani, S. Trillo, S. Carbone, and F. Montecchi, "Childhood autism rating scale (CARS) and autism behavior checklist (ABC) correspondence and conflicts with DSM-IV criteria in diagnosis of autism," *J. Autism Devlop. Disord.*, vol. 34, no. 6, pp. 703–708, 2004.
[18] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity," *J. Autism Devlop. Disord.*, vol. 37, no. 4, pp. 613–627, 2007.
[19] A. McCrimmon and K. Rostad, "Test review: Autism diagnostic observation schedule, second edition (ADOS-2) manual (Part II): Toddler module," *J. Psychoeducational Assessment*, vol. 32, no. 1, pp. 88–92, 2014.
[20] D. P. Wall, J. Kosmicki, T. Deluca, E. Harstad, and V. A. Fusaro, "Use of machine learning to shorten observation-based screening and diagnosis of autism," *Transl. Psychiatry*, vol. 2, no. 4, pp. e100–e100, 2012.
[21] J. Kosmicki, V. Sochat, M. Duda, and D. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Transl. Psychiatry*, vol. 5, no. 2, pp. e514–e514, 2015.
[22] J. N. Constantino et al., "Infant viewing of social scenes is under genetic control and is atypical in autism," *Nature*, vol. 547, no. 7663, pp. 340–344, 2017.
[23] T. Falck-Ytter, S. Bölte, and G. Gredebäck, "Eye tracking in early autism research," *J. Neurodevelopmental Disord.*, vol. 5, no. 1, pp. 1–13, 2013.
[24] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari, "Behavioral manifestations of autism in the first year of life," *Int. J. Devlop. Neurosci.*, vol. 23, no. 2, pp. 143–152, 2005.
[25] S. Ozonoff et al., "A prospective study of the emergence of early behavioral signs of autism," *J. Amer. Acad. Child Adolesc. Psychiatry*, vol. 49, no. 3, pp. 256–266.e2, 2010.
[26] L. Zwaigenbaum et al., "Early identification of autism spectrum disorder: Recommendations for practice and research," *Pediatrics*, vol. 136, no. Supplement_1, pp. S10–S40, 2015.
[27] G. Dawson et al., "Early social attention impairments in autism: Social orienting, joint attention, and attention to distress.," *Devlop. Psychol.*, vol. 40, no. 2, 2004, Art. no. 271.
[28] R. Zhang et al., "Towards a computer-assisted comprehensive evaluation of visual motor integration for children with autism spectrum disorder: A pilot study," *Interactive Learn. Environ.*, pp. 1–16, 2021. [Online]. Available: https://doi.org/10.1080/10494820.2021.1952273
[29] F. Zhu, S. Wang, W. Liu, H. Zhu, M. Li, and X. Zou, "Multi-modal machine learning system in early screening for toddlers with autism spectrum disorders based on response to name," *Front. Psychiatry*, vol. 14, 2023, Art. no. 34.
[30] F. Negin, B. Ozyer, S. Agahian, S. Kacdioglu, and G. T. Ozyer, "Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders," *Neurocomputing*, vol. 446, pp. 145–155, 2021.
[31] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Toward differential diagnosis of autism spectrum disorder using multimodal behavior descriptors and executive functions," *Comput. Speech Lang.*, vol. 56, pp. 17–35, 2019.
[32] C. Tang et al., "Automatic identification of high-risk autism spectrum disorder: A feasibility study using video and audio data under the still-face paradigm," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2401–2410, Nov. 2020.
[33] A. Di Martino et al., "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.
[34] A. Di Martino et al., "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," *Sci. Data*, vol. 4, no. 1, pp. 1–15, 2017.
[35] Y. Kong, J. Gao, Y. Xu, Y. Pan, J. Wang, and J. Liu, "Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier," *Neurocomputing*, vol. 324, pp. 63–68, 2019.
[36] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state fMRI using deep belief network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2847–2861, Jul. 2021.
[37] J. Zhou et al., "Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk," *Nature Genet.*, vol. 51, no. 6, pp. 973–980, 2019.
[38] N. Alotaibi and K. Maharatna, "Classification of autism spectrum disorder from EEG-based functional brain connectivity analysis," *Neural Computation*, vol. 33, no. 7, pp. 1914–1941, 2021.
[39] E. Abdulhay, M. Alafeef, H. Hadoush, and A. N, "Autism diagnosis via correlation between vectors of direct quadrature instantaneous frequency of EEG analytic normalized intrinsic mode functions," *Expert Syst.*, vol. 39, no. 3, 2022, Art. no. e12801.
[40] J. G. Frohna, "Failure to respond to name is indicator of possible autism spectrum disorder," *J. Pediatrics*, vol. 151, no. 3, pp. 327–328, 2007.
[41] N. Qiu et al., "Application of the still-face paradigm in early screening for high-risk autism spectrum disorder in infants and toddlers," *Front. Pediatrics*, vol. 8, 2020, Art. no. 290.
[42] L. Zwaigenbaum, S. Bryson, and N. Garon, "Early identification of autism spectrum disorders," *Behav. Brain Res.*, vol. 251, pp. 133–146, 2013.
[43] M. R. Talbott et al., "Brief report: Preliminary feasibility of the TEDI: A novel parent-administered telehealth assessment for autism spectrum disorder symptoms in the first year of life," *J. Autism Devlop. Disord.*, vol. 50, no. 9, pp. 3432–3439, 2020.
[44] A. Steiner, *Early Social-Communication Scales (ESCS)*, New York, NY, USA: Springer, 2013, pp. 1033–1034. [Online]. Available: https://doi.org/10.1007/978-1-4419-1698-3287
[45] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 17721–17732.
[46] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5202–5211.
[47] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.
[48] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2019, pp. 1487–1495.
[49] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.
[50] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[51] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2004.
[52] Tobii Pro AB, "Tobii pro lab," Computer software, Danderyd, Stockholm, 2014. [Online]. Available: http://www.tobiipro.com/
[53] X. Li, D. Zhang, M. Li, and D.-J. Lee, "Accurate head pose estimation using image rectification and a lightweight convolutional neural network," *IEEE Trans. Multimedia*, 2022.
[54] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.
[55] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall, "HOPE-Net: A graph-based model for hand-object pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6607–6616.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[57] O. Rodrigues, "Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire," *J. Math. Pures Appl.*, vol. 5, no. 380-400, 1840, Art. no. 5.
[58] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. UM-CS-2010–009, 2010.

[59] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life Images: Detection Alignment Recognit.*, 2008.
[60] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.
[61] M. Matilainen, P. Sangi, J. Holappa, and O. Silvén, "OUHANDS database for hand detection and pose recognition," in *Proc. 6th Int. Conf. Image Process. Theory Tools Appl.*, 2016, pp. 1–5.
[62] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4645–4653.
[63] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1949–1957.
[64] S. Narasimhaswamy, Z. Wei, Y. Wang, J. Zhang, and M. Hoai, "Contextual attention for hand detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9566–9575.
[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[66] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, First Quarter 2019.
[67] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, 2018.
[68] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2584–2593.
[69] The MPLab GENKI database, GENKI-4 K subset, 2009. [Online]. Available: http://mplab.ucsd.edu
[70] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.
[71] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via De-Albino and affinity," 2021, *arXiv:2103.10189*.
[72] M. M. Islam, S. Siddiqua, and J. Afnan, "Real time hand gesture recognition using different algorithms based on American sign language," in *Proc. IEEE Int. Conf. Imag. Vis. Pattern Recognit.*, 2017, pp. 1–6.
[73] J. D. Guerrero-Balaguera and W. J. Pérez-Holguín, "FPGA-based translation system from colombian sign language to text," *Dyna*, vol. 82, pp. 172–181, 2015.
[74] D. Deora and N. Bajaj, "Indian sign language recognition," in *Proc. 1st Int. Conf. Emerg. Technol. Trends Electron. Commun. Netw.*, 2012, pp. 1–5.
[75] P. P. Kumar, P. Vadakkepat, and A. P. Loh, "Hand posture and face recognition using a fuzzy-rough approach," *Int. J. Humanoid Robot.*, vol. 7, no. 3, pp. 331–356, 2010. [Online]. Available: https://doi.org/10.1142/S0219843610002180
[76] P. K. Pisharady, P. Vadakkepat, and A. L. Loh, "Attention based detection and recognition of hand postures against complex backgrounds," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 403–419, 2013. [Online]. Available: https://doi.org/10.1007/s11263-012-0560-52017.
[77] J. Triesch and C. von der Malsburg, "A system for person-independent hand posture recognition against complex backgrounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1449–1453, Dec. 2001.
[78] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.
[79] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.
[80] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[81] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.
[82] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011.
[83] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming mandarin ASR research into industrial scale," 2018, *arXiv:1808.10583*.
[84] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, 2017, pp. 1–5.
[85] V. I. Levenshtein et al., "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

Ming Cheng received the bachelor's degree in measuring and control technologies and instruments from China Jiliang University, and the master's degree in electrical and electronic engineering from the University of Hong Kong. He is currently working toward the PhD degree in computer science with Wuhan University. His research interests include speech signal processing and multimodal behavior analysis.



Yingying Zhang received the master's degree in medicine from Sun Yat-sen University, in 2022. She is currently a research assistant of Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include early diagnosis of autism, AI-assisted diagnosis and differential diagnosis of autism, and AI-assisted intervention of autism.

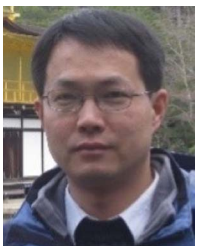
Yixiang Xie received the bachelor's degree in Chinese medicine pharmacy from Guangdong Pharmaceutical University. She is currently working toward the postgraduate degree in applied psychology with South China Normal University, Guangdong, China. Her research interests include multimodal behavior analysis for autism.

Yueran Pan received the bachelor's degree in statistics from Wuhan University, and the master's degree in data science with distinction from the London School of Economics and Political Science. She is currently working toward the PhD degree in computer science with Wuhan University. Her research interests include applications of multimodal behavior analysis to help children with autism.

Xiao Li received the bachelor's and master's degrees from Sun Yat-sen University, in 2018 and 2020, respectively. He is currently working toward the PhD degree with the University of Florida. His research interests include machine learning, computer vision, and 3D scene understanding.

Wenxing Liu received the master's degree in computer science from the Chongqing University of Technology. He is currently working toward the PhD degree in computer science with Wuhan University. His research interests include image processing, gaze estimation, and ASD behavior analysis.

Chengyan Yu received the bachelor's degree from the Taiyuan University of Technology, in 2021. He is currently working toward the postgraduate degree with the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include deep learning and facial expression recognition.



Dong Zhang received the BSEE and MS degrees from Nanjing University, in 1999 and 2003, respectively, and the PhD degree from Sun Yat-sen University, in 2009. He is currently an associate professor with the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, computer vision, affective computing, and information hiding.

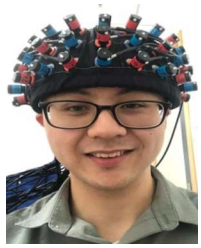
Yu Xing is currently working toward the PhD degree with Sun Yat-sen University. Currently, she works as a doctor of Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include the diagnosis and treatment of various developmental and behavioral disorders in children.

Xiaoqian Huang received the bachelor's degree from the Guangzhou University of Chinese Medicine, in 2014. She is currently a research assistant with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include the early development of infant and toddler ability and AI-assisted intervention of autism.

Fang Wang received the master's degree in medicine from Sun Yat-sen University. She is currently a doctor of Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include the diagnosis and treatment of various developmental and behavioral disorders in children.

Cong You received the master's degree in medicine from Sun Yat-sen University. She is currently a doctor of Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include the diagnosis and treatment of various developmental and behavioral disorders in children.

Yuanyuan Zou is currently the head therapist with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. She has more than 20 years of clinical experience in the assessment, intervention, and family guidance of children with developmental disabilities.



Yuchong Liu received the MD degree in pediatrics from Sun Yat-Sen University, in 2021. He is currently working toward the PhD degree in pediatrics with Sun Yat-sen University. He is particularly interested in understanding functioning and well-being in autism, developing, evaluating, and translating evidence-based socio-emotional and strengths-based interventions into practice.

Fengjing Liang is currently working toward the PhD degree with Sun Yat-sen University. Currently, she works as a doctor of Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include the diagnosis and treatment of autism spectrum disorder and other children's developmental diseases.

Huilin Zhu received the PhD degree in psychology from South China Normal University. She is currently a research fellow with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include neuropsychology, developmental psychology, and clinical psychology.

Chun Tang is currently an Associate Professor with the Department of Pediatrics, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include developmental-behavioral pediatrics and the diagnosis and treatment of autism and other developmental disabilities.

Hongzhu Deng received the PhD degree in medicine from Sun Yat-sen University, in 2010. She is currently an associate professor with the Department of Pediatrics and the director with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Her research interests include the diagnosis and treatment of autism and other developmental disabilities.

Xiaobing Zou is the academic leader with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. He has rich clinical experience in the field of developmental behavioral disorders for children. His research focuses on early diagnosis and behavioral intervention strategies for children with the spectrum and other developmental disabilities.

Ming Li received the PhD degree in electrical engineering from the University of Southern California, in 2013. He is currently an associate professor of electrical and computer engineering with Duke Kunshan University. He is also an adjunct professor with the School of Computer Science, Wuhan University. His research interests include audio, speech, language, and multimodal behavior signal analysis and interpretation.