Research on a Multi-source Perception-fused Intelligent Intervention Robot System for Children with Autism Spectrum Disorder

Yinuo Qiu[1,*], Xiaolei Pan[1], Chenxi Jiang[2], Minzhu Liu[1]
[1]School of Psychology, Zhejiang Normal University, Jinhua 321004, China
[2]College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua 321004, China
*Corresponding email: 2175302064@qq.com

Abstract

Addressing the pressing challenges in autism rehabilitation - such as teacher shortages, homogenized intervention approaches, and high costs of overseas equipment - this study proposes and implements an intelligent intervention robot system that integrates multi-source sensory data. First, the study constructed China's first facial emotion dataset for children with autism spectrum disorder facial expression dataset (ASDface). Based on this dataset, it designed an ASDface-AlexNet recognition model incorporating a dual-attention mechanism. This model achieved an average recognition accuracy of 89.31% even for subtle facial expressions. Building upon this, the study further integrated four types of signals - facial, ocular, head movement, and task-response - to develop a cognitive-load assessment model. This enables the robot to dynamically and adaptively adjust intervention strategies in real time. The system comprises integrated robotic hardware and a collaborative mobile app for home-institution coordination, forming a closed-loop intervention framework. An eight-week controlled trial demonstrated that this system significantly reduced the proportion of cognitive overload during children's training sessions (11.30% in the experimental group vs. 42.70% in the control group). It also effectively enhanced children's core competencies and greatly improved the sustainability of home-based interventions. This research provides a localized technological solution to enhance the accessibility and inclusivity of autism intervention services in China.

Keywords

Autism spectrum disorder, Multi-source sensory fusion, Facial emotion recognition, Cognitive load, Intelligent intervention robot

Introduction

Autism spectrum disorder is a heterogeneous, pervasive neurodevelopmental disorder whose core clinical manifestations include deficits in social communication, restricted and narrow interests, and repetitive, stereotyped behaviors. Most affected children exhibit varying degrees of sensory processing dysfunction, differences in cognitive development, and emotional regulation difficulties.

China's series of documents on digitalization of special education and technology-assisted disability support have established a top-down policy framework. This framework provides robust support for the research and implementation of intelligent rehabilitation devices for autism. Among these documents, the Outline for Building a Powerful Education Country (2024-2035) explicitly sets forth the goal of promoting inclusive special education and encourages the development of lightweight, digital assistive tools tailored to individuals with developmental disorders. The 14th Five-Year Action Plan for Promoting the Development of Special Education identifies the construction of smart special education campuses as a key task. It advocates the use of intelligent sensing technologies to deliver tiered, adaptive training and thereby address the uneven distribution of rehabilitation resources across regions. The Implementation Plan for the Autism Children Care and Support Initiative (2024-2028) directly positions the development of intelligent rehabilitation equipment as a central pathway for expanding the supply of rehabilitation services.

Zhejiang Province and Jinhua City have simultaneously introduced local supplementary implementation guidelines, launching a free screening program for children with autism aged 0-6 years old. They have also opened local children's rehabilitation centers and welfare institutions as sites for scientific research and empirical testing, providing practical platforms for sample collection and pilot testing of robotic devices. The algorithmic architecture, hardware systems, and service models designed in this study are not limited by geographic boundaries and possess universal value for promotion across special education institutions at all levels nationwide.

Under the guidance of government policies, the autism rehabilitation industry continues to face a long-term structural imbalance between supply and demand. According to publicly released data from the 2024 Statistical Bulletin on the Development of China's Disability Sector, there are a total of 4,930 institutions nationwide that hold formal qualifications for autism intervention. Yet the number of professionally trained rehabilitation practitioners currently employed stands at only 58,874. These figures align with typical patterns observed in developing countries. A 1:10 calculation of rehabilitation service ratios indicates that the current institutional capacity can accommodate approximately 600,000 child patients. Meanwhile, the number of autistic children under the age of 14 in China has reached between 3 million and 5 million, with nearly 200,000 new cases diagnosed each year on average. The uneven distribution of resources further widens the intervention gap: more than 70.00% of children aged 0 to 6 - the critical early intervention period - receive less than 12 hours of institutional training per week. Moreover, there is a particularly severe shortage of certified rehabilitation therapists in county-level and township areas.

Moreover, currently, domestic rehabilitation institutions generally rely on paper-based assessment forms to conduct phased evaluations. The conclusions drawn from these assessments heavily depend on the individual rehabilitation therapists' professional experience, lacking dynamic monitoring indicators that can be continuously collected and quantitatively analyzed throughout the training process. Meanwhile, industry survey data reveal that only 45.40% of frontline rehabilitation teachers have received comprehensive, systematic training in special education. Half of all practitioners enter the field directly after undergoing short-term, rapid training courses, with little awareness of how to tailor interventions to children's cognitive development levels and individual personalities. As a result, the "one-size-fits-all" approach to training has become a major factor limiting the effectiveness of rehabilitation efforts. Thus, it is evident that the traditional manual intervention model inherently suffers from unavoidable shortcomings that make it difficult to meet the differentiated rehabilitation needs of children with autism. Accelerating the research and development of intelligent products for autism is one viable solution to address this need.

Literature review

Current research status of robot-assisted interventions for autism outside China

Research on robot-assisted autism rehabilitation started earlier abroad and has already developed a relatively complete theoretical and practical framework. Several controlled experimental studies have confirmed that humanoid robots' interactive approach is more appealing to children with autism than human therapists. With features such as highly controllable interaction behaviors, repeatable training procedures, and lower social pressure, these robots can effectively reduce children's tendency toward social avoidance. They also enhance core social skills - including eye contact, action imitation, and joint attention. In the field of multimodal perception architecture research, the Adaptive Robotic Intervention Architecture (ARIA) has been proposed, integrating visual motion capture and speech signal acquisition to achieve real-time tracking of children's interactive behaviors. However, the system's emotion recognition module, trained on datasets from the general population, suffers from significant shortcomings in accurately detecting subtle emotions in children with autism [1]. A systematic review of clinical applications of autism intervention robots found that mainstream overseas devices generally lack mechanisms for dynamically adjusting training content based on children's real-time states; most products adopt fixed training content and pacing [2]. In affective computing research, Heinsfeld leveraged the Autism Brain Imaging Data Exchange (ABIDE) brain imaging dataset to develop an autism-assisted screening tool. Yet this study focused solely on medical diagnostic scenarios and did not address the need for real-time adjustments in rehabilitation training [3]. The NAO robot - the most widely used device - comes equipped only with basic unimodal facial expression recognition capabilities, making it incapable of capturing subtle avoidance behaviors such as gaze shifts or head turns. Moreover, it lacks a cognitive load assessment dimension and does not possess adaptive intervention adjustment capabilities. Recent synthesis evidence underscores both the promise and the limitations of this field: a systematic review and meta-analysis reported benefits across social-communication outcomes. However, it also emphasized substantial heterogeneity in robot platforms, intervention dose, and outcome measures [4]. A systematic review of randomized controlled trials similarly concluded that the clinical readiness of social robots remains constrained by limited sample sizes and variable methodological quality [5].

Meanwhile, although humanoid robots such as NAO, Qtrobot, and Milo dominate the global market, their individual unit prices generally exceed 10,000 yuan, far beyond the affordability of most Chinese households. Moreover, the training datasets used for these robots are primarily drawn from children in Europe and the United States. Their built-in training programs are designed based on Western special education systems. This makes them ill-suited to effectively address the emotional expression characteristics of Chinese children with autism, which are characterized by subtle facial expressions and frequent avoidance behaviors.

Furthermore, in terms of the datasets employed, widely used public face datasets in the field of affective computing, such as Facial Expression Recognition 2013 (FER2013) and Cohn-Kanade Plus (CK+), are both collected from typically developing adults and children. This results in significant mismatches when applied to the emotional expression patterns of children with autism. Children with autism exhibit even fewer facial expressions and weaker emotional cues; during training, they frequently display avoidance behaviors like shifting gaze or turning their heads. The recognition accuracy of emotional changes in eyebrows, eyes, and facial contours is significantly lower in autistic children compared to neurotypical children. Consequently, recognition models trained on generic datasets struggle to capture these sparse and nuanced emotional features, making it highly likely that interventions will misinterpret the children's true psychological states.

Currently, neither the academic nor the industrial communities in China have developed any specialized solutions tailored specifically for this population. A standardized, publicly available facial dataset for children with autism spectrum disorder, featuring synchronized annotations of emotional and cognitive load correlations, is urgently needed. Data shortages have become the core bottleneck hindering the development of specialized emotion recognition devices for autism.

Current status of autism robot research in China

Research on intelligent interventions for autism in China is still in the preliminary exploration stage. Most of the existing findings suffer from a disconnect between technological development and rehabilitation practice, showing a research bias toward algorithmic sophistication at the expense of practical implementation. At the hardware development level, some researchers have built an interactive robotic simulation system for autism. However, this system only provides basic picture-exchange communication training functionality and lacks a multi-source real-time behavioral sensing module. Researchers have combined the Picture Exchange Communication System (PECS) with a humanoid robot to demonstrate the robot's auxiliary role in social skills training. However, the device cannot autonomously recognize children's facial expressions, and the children's emotional states still rely on manual data entry. On the level of single-emotion recognition products, a facial emotion-training tool for children with autism was developed, with its interactive design refined through questionnaire surveys and biofeedback experiments. Yet, this product supports only unimodal face recognition and fails to integrate multiple behavioral indicators such as head avoidance and task responsiveness, thus lacking a comprehensive, hierarchically structured intervention task library.

In the field of theoretical review, previous studies have systematically summarized the research findings on the application of social robots in autism rehabilitation. Robots can enhance children's behavioral performances in areas such as imitation and proactive social interaction. At the same time, several shortcomings in domestic research have been identified. These include the scarcity of autism-specific samples, the lack of lightweight algorithms capable of accurately recognizing subtle facial expressions, and the inadequacy of a well-coordinated support system involving both families and institutions. Other literature has reviewed studies related to contextual emotional understanding in autism. Directions have been proposed for the development of standardized specialized training materials and multimodal real-time monitoring devices. These provide theoretical reference for the design of the multi-source perception fusion architecture in this study.

Four core research gaps can be identified through a comprehensive review of existing research findings both domestically and internationally.

First, there is a gap in data resources. China lacks a standardized facial dataset for children with autism that is annotated with labels linking emotions to cognitive load. General population datasets exhibit poor adaptability. Second, there is a gap in algorithmic technology. Traditional convolutional networks struggle to effectively focus on the sparse facial emotional features of children with autism. Emotion-computing models rely on manual trial-and-error for parameter tuning, resulting in low research and development (R&D) efficiency. Third, there is a gap in evaluation systems. Current studies rely solely on single facial-emotion assessments to gauge children's training status, failing to integrate multi-dimensional indicators such as avoidance behaviors and task-response delays to quantify cognitive load. Fourth, there is a gap in intervention systems. Overseas robots are expensive and lack adequate local adaptation. Domestically available general-purpose children's robots lack specialized special education modules, and there is a shortage of integrated closed-loop intervention systems that combine robotic hardware, perception algorithms, and home-based assistive terminals.

The existing research clearly reveals significant innovation gaps, providing ample room for this study to explore innovations across four dimensions: datasets, optimized algorithms, multi-source evaluation, and collaborative intervention systems.

Theoretical foundation

Cognitive Load Theory serves as the core theoretical foundation for the lightweight real-time evaluation model used in this study. The theory posits that human working memory has a fixed capacity limit. When task difficulty exceeds this capacity or sensory stimuli become overly complex, cognitive overload may be triggered, leading to negative behaviors such as gaze avoidance, emotional irritation, and task resistance. For tasks that are too easy, cognitive load becomes insufficient, making it impossible to achieve effective skill enhancement [6]. Children with autism have weaker cognitive information-processing abilities than typically developing children, and they are more prone to experiencing imbalances in cognitive load during training. In this study, cognitive load is taken as the core dimension for assessing a child's teachability. Facial microexpression intensity, blink frequency, duration of gaze avoidance, and task error rate are selected as external, objective indicators of cognitive load. Based on these indicators, three levels of training states - cognitive overload, appropriately balanced load, and cognitive underload - are defined, accompanied by corresponding intervention and adjustment rules. By establishing a logical bridge between multi-source perceptual signal acquisition and adaptive intervention strategy adjustments, cognitive load theory provides the underlying theoretical foundation for the robot's intelligent decision-making module.

Sensory integration theory provides the underlying logical foundation for designing the intervention task library in this study. According to this theory, an individual's adaptive social behavior relies on the integration of multisensory information from vision, hearing, touch, vestibular sense, and proprioception. Children with autism spectrum disorder commonly exhibit sensory integration dysfunction, showing either hypersensitivity or hypoesthesia toward specific sensory stimuli, which directly impairs the development of fine motor skills and social communication abilities [7]. In this study, hierarchical training tasks have been designed - covering visual matching, auditory discrimination, balance and coordination, and fine hand manipulation - based on the perceptual characteristics of children with autism. The robot's multimodal sensing system can autonomously adjust the intensity and duration of visual and auditory stimuli in real time according to the child's cognitive load. This avoids sensory overload stimuli that easily trigger frustration. This approach enables a gradual, moderately intensive multisensory rehabilitation training program that organically integrates traditional sensory integration interventions with artificial intelligence (AI)-powered real-time perception technology.

The Natural Developmental Behavioral Intervention Theory serves as the core guiding principle for robotic interaction logic and training task design. This theory emphasizes that autism intervention should follow the natural developmental patterns of children, adhering to four key principles: early intervention, personalized tiered approaches, gamified presentation, and positive reinforcement. It rejects rigid, rote-learning training methods [8]. In this study, the theory's core tenets have been fully integrated into the robot's entire interactive process. All training tasks are presented in the form of engaging games; upon successfully completing a task, the robot simultaneously provides verbal and visual positive reinforcement. The task library is organized into graded levels based on children's age and skill levels, and the system automatically matches personalized task sequences by leveraging real-time sensory data. The robot employs a gentle speaking pace and a low-pressure, visually intuitive interaction mode that aligns with the social preferences of children with autism. This addresses the stiff and mechanical shortcomings inherent in traditional smart-device interactions. There is evidence from a systematic review of the Joint Attention, Symbolic Play, Engagement, and Regulation (JASPER) intervention, which targets joint attention and play within naturalistic developmental contexts. The present task design prioritizes responsive interaction and contingent reinforcement [9].

Scholars from home and abroad, building on the three foundational theoretical frameworks mentioned above, have gradually developed a multifaceted theoretical framework for "technology-enabled rehabilitation". It clearly highlights the unique advantages of humanoid robots as assistive intervention tools for autism. Compared to human rehabilitation therapists, robotic interactions are highly standardized, free from emotional fluctuations, and allow training content to be repeated infinitely, thereby reducing anxiety in autistic children during social learning. Human-machine interaction technologies - such as facial emotion recognition - can objectively and quantitatively collect behavioral data throughout the entire process. This overcomes the limitations of subjective bias inherent in manual assessments and provides data support for the development of personalized intervention plans. This study integrates the core tenets of the multifaceted theoretical framework, using intelligent robots as technological carriers to achieve the practical integration of psychological rehabilitation theories with computer vision technologies.

Definition of core concepts

This study selected children aged 5-12 years old with mild to moderate autism as the entire subject population. Multi-source perceptual fusion refers to the simultaneous acquisition of heterogeneous behavioral signals, including facial microexpressions, eye gaze patterns, head avoidance behaviors, and task responses. These signals are acquired using multimodal sensing devices such as high-definition cameras and three-dimensional (3D) laser rangefinders. These signals are then comprehensively integrated to generate a holistic, quantified assessment of children's psychological states, distinguishing it from single-modal emotion recognition approaches that rely solely on facial images.

Cognitive load and engagement are the two core dimensions for measuring the "teachability" of a child's training process. Cognitive load reflects the degree to which the task difficulty matches the child's cognitive processing capacity, while engagement represents the child's proactive willingness to interact. These two factors are comprehensively assessed through multi-source perceptual signals and are collectively categorized into three levels: cognitive overload, appropriate load, and cognitive underload. The ASDface-AlexNet attention-optimized model is a lightweight recognition network built upon the AlexNet convolutional neural network framework, incorporating both channel attention and spatial attention modules. It is specifically optimized for subtle facial emotional features in children with autism. Compared to general deep learning models, it demonstrates a significant improvement in accuracy when detecting low-intensity microexpressions and proactive avoidance behaviors.

The family-institution collaborative closed-loop intervention system uses intelligent intervention robots as data collection devices. This enables real-time synchronization of children's emotional states and developmental progress data between the rehabilitation therapist's terminal and the parent's mobile app mini-program. This provides a unified service model. Institutions are responsible for developing professional, stage-based training plans, while families carry out daily, ongoing supplementary training. Relying on long-term, bidirectional data accumulation, the system continuously optimizes intervention tasks in both settings, creating a closed-loop rehabilitation model characterized by two-way feedback and sustained, long-term effectiveness.

Finally, regarding quantification of intervention effects in controlled interventions, field research captured genuine needs from frontline rehabilitation settings. Three children's rehabilitation and education centers in Jinhua were visited, and the implementation process of conventional interventions was documented. The functional limitations of existing training equipment were also analyzed, and in-depth discussions were conducted with institutional administrators and practicing rehabilitation therapists to gather first-hand industry demand information.

At the same time, questionnaires and semi-structured interviews were used to identify the key pain points faced by families of children with autism. Both paper-based and customized electronic questionnaires were distributed to parents of affected children, totaling 81 distributed questionnaires and recovering 73 valid responses. Reliability and validity tests were conducted on the questionnaires using Cronbach's alpha coefficient and Kaiser-Meyer-Olkin (KMO) value. In addition, a total of ten in-depth interviews were conducted with five families of children with autism, producing over 50,000 words of detailed interview transcripts. From these transcripts, families' core functional requirements and usage expectations for intelligent rehabilitation devices were extracted.

To construct a quantitative evaluation system for intervention effects, the Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation Method were utilized. Core sensory integration training tasks that are well-suited for children with autism were then selected. An evaluation framework encompassing six key dimensions was also developed. These dimensions include joint attention, mental theory, visual matching, fine motor skills, cognitive comprehension, and concept categorization. This framework ultimately resulted in standardized quantitative indicators for evaluating intervention effectiveness.

Model design and hardware

Construction of a facial emotion dataset specifically for children with autism

Dataset construction is an upstream task in model design. For the data sources of our dataset, our team relied on clinical diagnostic records from rehabilitation centers to select 43 children with mild to moderate autism spectrum disorder. The sample includes 31 boys and 12 girls, aged between 5 and 12 years old. All participating children were able to independently complete basic human-computer interaction training tasks without experiencing severe sensory overstimulation or being unable to adapt to camera filming. All guardians of the participating children signed written informed consent forms for data collection. All image data have been de-identified and will be used solely for academic model training, with strict adherence to child privacy protection regulations.

The study employed a scenario-based emotional induction method, collecting image data in two real-world settings: indoor rehabilitation training and outdoor free play. This ensured that the dataset comprehensively covers the authentic, everyday rehabilitation scenarios experienced by children. In the indoor training setting, standardized sensory integration tasks such as visual matching, puzzle solving, and picture recognition were implemented. By adjusting task difficulty and providing positive or negative feedback, children were induced to experience five core emotions: happiness, calmness, boredom, frustration, and surprise. Simultaneously, an OpenMV4 high-definition camera was used continuously to capture frontal facial images, recording the corresponding cognitive load levels of each task. In the outdoor activity setting, natural facial images were collected during children's free play, simple social interactions, and environmental exploration. At the same time, active avoidance behaviors - such as head turning, gaze shifting, and body retreating - were recorded. Combined with on-site observations by rehabilitation therapists, each frame of imagery was assigned the appropriate emotional label.

After accumulating 236 minutes of continuous data collection, the research team obtained 1,874 raw image samples. Following a series of preprocessing steps - including image screening, noise filtering, brightness adjustment, and minor angle correction - eventually 3,245 standardized, valid images were generated. These images were uniformly labeled with five emotion categories, thus creating a comprehensive facial emotion dataset specifically tailored for children with autism.

The annotation of the original images was jointly carried out by two researchers with specialized backgrounds in special education. Prior to starting the annotation, both researchers underwent specialized training on emotion recognition in children with autism, and a unified set of criteria for judgment was established. Happiness is indicated by a relaxed facial expression, upward-turned corners of the mouth, active eye contact, and accompanying joyful body movements. Calmness is characterized by a face that shows no obvious facial expressions, stable performance of basic tasks, and the absence of avoidance or irritability. Boredom is identified by frequent shifting of gaze away, head lowering, and repetitive, stereotyped hand movements. Anger is marked by tightly furrowed brows, a tense facial expression, and active interruption of the task. Frustration is indicated by briefly widened eyes, slight tension in the facial muscles, and brief but focused attention toward unfamiliar stimuli.

After two researchers independently completed the annotation of all samples, a joint review of samples with inconsistent annotations was conducted. The labels were refined by integrating video recordings collected at the same time and on-site observations made by rehabilitation therapists. Ultimately, the inter-rater agreement coefficient exceeded 0.85.

The original images suffer from issues such as skewed shooting angles, uneven lighting, image blurriness, and interference from irrelevant background elements. The study implements a four-stage standardized preprocessing pipeline in sequence. The first stage involves face cropping: leveraging the Multi-Task Cascaded Convolutional Networks (MTC) algorithm to locate the facial region, remove redundant background areas, and standardize the image dimensions. The second stage focuses on noise filtering - directly discarding invalid samples that are blurry, heavily occluded, or lack valid facial features. The third stage is data augmentation: expanding the total sample size through horizontal flipping, subtle brightness adjustments, slight rotations, and Gaussian blurring, thereby reducing the risk of model overfitting. The fourth stage is pixel normalization: standardizing the image's pixel range and dimension specifications to align with the input requirements of the subsequent ASDface-AlexNet model.

ASDface-AlexNet recognition model design

The traditional AlexNet network was trained using facial data from the general population. During the convolutional feature-extraction process, it paid insufficient attention to the sparse and subtle emotional features of children with autism. This resulted in significant errors in recognizing low-intensity emotions such as boredom and surprise. To address this limitation, this study embeds both a channel-attention module and a spatial-attention module into the original AlexNet architecture, thereby constructing an optimized ASDface-AlexNet recognition model. The spatial-attention module automatically focuses on key facial regions associated with emotions, such as the eyebrows, eyes, and corners of the mouth. The channel-attention module selectively filters feature channels that carry high discriminative value. This dual mechanism works synergistically to enhance the extraction of subtle microexpressions. Meanwhile, a hierarchical hyperparameter auto-search algorithm is introduced to tackle the issues of low efficiency in manual parameter tuning and weak generalization ability inherent in emotion-computing models. This simultaneously improves both the model's recognition accuracy and its real-time local inference performance.

The model consists of five convolutional pooling modules, a dual-channel attention embedding module, three fully connected layers, and a Softmax classification output layer. The shallow convolutional layers extract facial contour and basic texture features, paired with max-pooling layers to reduce the size of the feature maps. After the middle convolutional layers, a spatial-and-channel dual-attention module is cascaded to assign weighted values to the feature maps. This module enhances the weights of features in emotion-critical regions while simultaneously suppressing interference from irrelevant features such as background and skin texture. The deep convolutional layers perform multi-scale, deep fusion on the weighted, high-value emotion features. The fully connected layers map the fused features onto five emotional dimensions, and the Softmax function outputs the probability scores corresponding to each emotion, thereby completing the emotion classification and identification.

The model training environment was set up on the Ubuntu 16.04 operating system and trained using the PyTorch deep learning framework. The hardware is equipped with an NVIDIA GeForce GTX 2080 graphics card. The optimizer chosen was Adam, with an initial learning rate of 1e-4. A cosine annealing strategy was adopted to dynamically reduce the learning rate during training. The model underwent 70 epochs of iterative training, with each batch size set at 32. The MultiLabelSoftMarginLoss function was selected as the loss function to perform backpropagation and adjust the error.

Traditional parameter tuning methods for emotion computing models, such as grid search and random search, suffer from drawbacks: high randomness, time-consuming iterations, and difficulty pinpointing the optimal parameters. In this study, a hierarchical automated search algorithm is adopted to optimize model hyperparameters, which consists of two stages: coarse search and fine search. During the coarse search stage, the search ranges for learning rate, number of convolutional kernels, and attention weight coefficients are expanded, and multiple parameter combinations are trained in batches. Simultaneously, the validation and test loss data for each combination are recorded. In the fine search stage, regression trees are leveraged to analyze the loss data obtained from the coarse search, and the optimal parameter ranges are narrowed down. The search is then iteratively refined within a limited parameter space, and the optimal hyperparameter combination that best suits the task of emotion recognition for children with autism is ultimately identified. This approach significantly reduces the time and cost associated with manual parameter tuning while simultaneously enhancing the model's generalization capability in recognizing subtle emotional expressions.

Construction of a cognitive load assessment model for multi-source perception fusion

Relying solely on single facial emotion indicators is insufficient to fully assess the true training status of children with autism; some affected children exhibit a disconnect between their facial emotional expressions and their internal cognitive load. Combining the core tenets of cognitive load theory with the practical experience of frontline rehabilitation therapists, this study has identified and screened four categories of objective behavioral indicators. These indicators can be collected in real time via sensors. This approach thereby establishes a multi-dimensional integrated assessment system.

The first category comprises facial microexpression indicators, including emotion classification results, the intensity duration of expressions, and the frequency of expression transitions. The second category involves eye-gaze indicators, covering the duration of each valid gaze, the total duration of gaze avoidance, and the average blink rate. The third category pertains to active avoidance behaviors, encompassing head-turning amplitude, body-tilt angle, and the frequency of head turns. The fourth category consists of task-response indicators, including task-completion accuracy, the latency of individual thought responses, and the number of consecutive errors.

All four categories of indicators are synchronously collected through the robot's built-in sensing devices and uniformly standardized into a 0-1 numerical range, thereby eliminating computational biases caused by differing measurement units. Eye-tracking research has shown that gaze-derived features are valuable for distinguishing autistic from typically developing children in conversational contexts [10]. Multimodal recordings of eye movements and facial mimicry have also been applied to examine automatic emotion processing in autistic boys [11]. These findings support the inclusion of gaze and avoidance measures alongside facial-expression features.

The study, based on the Analytic Hierarchy Process, assigns differentiated weights to four categories of multi-source perceptual indicators and constructs a comprehensive cognitive load calculation formula. Three categories of children's training states are defined according to the overall score. Cognitive overload corresponds to an overall score above a high threshold. Externally, this manifests as prolonged gaze avoidance, frequent head turns, consecutive task errors, and a disproportionately high proportion of negative emotions. The system determines that the current task difficulty exceeds the child's cognitive capacity and initiates de-escalation adjustment strategies. Optimal load corresponds to an overall score within the intermediate range. In this case, the child can maintain stable eye contact, demonstrates good task completion, and exhibits a balanced distribution of positive and negative emotions. The system maintains the original training pace and content. Cognitive underload corresponds to an overall score below a low threshold. Throughout the task, there is no obvious avoidance behavior, zero task errors, and no significant emotional fluctuations. The system determines that the task difficulty is too low and automatically increases the complexity of the training and introduces more multisensory interactive stimuli.

After the assessment model outputs the child's cognitive load level, the robot simultaneously executes a corresponding intervention and adjustment strategy, thereby forming a complete closed-loop control logic. In a state of cognitive overload, the system automatically lowers the current task level, shortens the duration of each training session, and inserts animated break intervals. It also reduces the intensity of audiovisual stimuli and switches to simple, low-pressure interactive games. When the cognitive load is appropriate, the system maintains the task difficulty, training duration, and audiovisual stimulus parameters unchanged. Once the child completes the target task, the system automatically delivers positive reinforcement content. In a state of cognitive underload, the system raises the task level, introduces multisensory assistive tools, and extends the duration of each training session. It also adds scenario-expanding questions to increase the challenge of the training. The entire assessment model is deployed in a lightweight manner on the robot's embedded main control board, eliminating the need for cloud-based high-performance computing resources. All metric calculations and strategy outputs can be performed locally, making it well-suited for application scenarios in home settings and small rehabilitation facilities that lack specialized computing equipment.

Robot hardware system and collaborative mini-program development

The iterative hardware design of two generations of humanoid robots has been completed. The main body of the robot is made from lightweight 304 steel, with a total weight of 4.12 kg and a motion response speed of 6 m/s. The hardware system is divided into three major modules: the perception layer, the master control and drive layer, and the interaction output layer. The perception layer is equipped with a 7-inch touch-screen display and an OpenMV4 wide-angle camera for capturing facial images. It also features an SC1 3D laser sensor for detecting environmental obstacles and tracking children's body postures, as well as a multi-array microphone array for simultaneously collecting voice interaction signals. The master control and drive layer is centered around an Arduino control board, integrating motor drivers, wireless communication modules, and power-monitoring submodules. An embedded Linux platform hosts a lightweight ASDface-AlexNet recognition model, enabling on-device image inference and computation. The robot's chassis is fitted with Cartridge Direct-Drive Rotary (CDDR) servo torque motors, enabling autonomous mobility, steering, and close-range following functions. The interaction output layer uses the touch screen to display training animations and instructional images, and includes a built-in audio module that outputs smooth prompts and engaging sound effects. The robot's onboard lights change in synchronization with task completion status, creating a low-pressure, multisensory interaction environment. The procurement cost of the hardware kit is controllable, offering a significant price advantage over overseas devices such as Milo and Qtrobot, which typically cost tens of thousands of yuan. This makes our robot well-suited for large-scale deployment in grassroots rehabilitation centers and ordinary households.

At the same time, a companion mobile mini-program has been designed that establishes a two-way data channel between rehabilitation therapists and parents of child patients, creating a closed-loop intervention system connecting home and institutional settings. The overall functionality is divided into three core modules.

The child individual assessment module synchronously integrates all multi-source sensory data collected by the robot, fitting growth curves for six key dimensions. These dimensions include joint attention, theory of mind, visual matching, fine motor skills, cognitive comprehension, and conceptual categorization. The curves quantitatively illustrate the long-term trends in a child's developmental progress. The personalized intervention plan module automatically stratifies and delivers training games, animated tutorials, and simple home-based sensory integration programs tailored to each child's real-time cognitive load and overall competency level. Rehabilitation therapists can remotely adjust task difficulty and training frequency manually and issue specialized intervention guidelines. The data sharing and feedback module enables parents to view in real time their child's daily training duration, emotional fluctuations, and task completion status. On the therapist side, it allows for batch viewing of intervention data from all participating children and generates periodic stage-specific rehabilitation assessment reports. Both parties can submit feedback on device usage and training adaptation, supporting continuous system iteration and optimization.

The mini-program employs the Message Queuing Telemetry Transport (MQTT) communication protocol to achieve real-time data synchronization between devices and the mobile end. The backend is equipped with a lightweight cloud storage service and incorporates basic data protection mechanisms to mitigate the risk of children's private data being leaked.

Methods

The study selected 24 children aged 5-12 years old with mild to moderate autism from cooperative rehabilitation centers. The participants were randomly assigned to an experimental group and a control group, with 12 participants in each group. A controlled intervention experiment lasting eight weeks was conducted.

The comparative study consisted of two modules: an algorithm accuracy comparison and an intervention effectiveness comparison. For the algorithm experiment, a training environment was built using the PyTorch deep learning framework, and under a unified, self-built dataset, the recognition accuracy of the original AlexNet model was compared with that of the ASDface-AlexNet model.

In the intervention experiment, a robot-assisted experimental group and a traditional manual training control group were established, and a controlled study lasting eight weeks was carried out. The experimental group used our robot to conduct daily 30-minute, tiered, gamified training sessions, with the device adjusting intervention strategies in real time based on multi-source sensory data. Parents simultaneously used a companion app to continue the training at home. The control group followed a traditional, standardized sensory integration training regimen, receiving fixed three-session weekly training sessions at the facility, without any standardized home-based guidance or support.

The study included three observation points: before the start of the experiment, at the end of the fourth week, and at the end of the eighth week. At each of these points, three certified rehabilitation therapists assessed three core indicators - joint attention, emotion recognition, and fine motor skills - using a unified assessment scale. Simultaneously, experimental data were recorded using the National Aeronautics and Space Administration Task Load Index (NASA-TLX) subjective cognitive workload scale, objective task completion accuracy, and duration of active eye contact. This enabled a comprehensive evaluation of the intervention's effectiveness. To facilitate replication, the intervention was described with attention to core Template for Intervention Description and Replication (TIDieR) elements, including materials, procedures, dose, and tailoring [12]. Given the AI-enabled decision component, the reporting framework also specifies the algorithmic inputs and adaptive logic in line with Consolidated Standards of Reporting Trials - Artificial Intelligence (CONSORT-AI) considerations for AI interventions [13].

Results

Results of the model control experiment

The study used a self-built facial dataset of children with autism as a unified test sample to conduct a comparative experiment on the recognition accuracy of the original AlexNet model versus the ASDface-AlexNet model. The experimental data are shown in Table 1.

Table 1. The results of the model control experiment.

Model	Overall validation set accuracy%	Happy%	Calm%	Boredom%	Frustration%	Surprise%
Original AlexNet	85.73	94.00	81.00	87.00	92.00	82.00
ASDface-AlexNet	89.31	95.23	83.07	91.56	93.26	85.31

Based on the comprehensive analysis of experimental data, the ASDface-AlexNet model achieved an average emotion recognition accuracy of 89.31%, representing a 3.58-percentage-point improvement over the original AlexNet. Among these, the most challenging subtle emotions - boredom and surprise - showed the most significant improvements in recognition accuracy: boredom recognition accuracy improved by 4.56%, and surprise recognition accuracy improved by 3.31%. This fully validates the optimizing effect of the dual-channel attention mechanism on the recognition of low-intensity microexpressions in children with autism. The model's local single-inference latency is kept within the hundred-millisecond range, enabling it to meet the operational requirements for real-time dynamic adjustment of training strategies during robotic intervention.

Results of the control intervention experiment

After the completion of the eight-week comprehensive intervention cycle, multidimensional observational data revealed significant intergroup differences. At the level of core competency indicators, the experimental group's children showed significantly greater improvements. These improvements include emotional recognition accuracy, duration of sustained joint attention, and fine-motor skill completion rates compared to the control group. In the control group, which followed a standardized training protocol, individual differences in children's ability gains were substantial. Some participants consistently exhibited resistance to training and declining engagement. By contrast, the experimental group, leveraging a dynamic adaptive regulation mechanism, maintained stable engagement levels across all participating children.

In terms of cognitive load data, the proportion of observations indicating cognitive overload in the latter half of training reached 42.70% in the control group, with frequent occurrences of negative behaviors such as avoidance and irritability. In the experimental group, only 11.30% of observations indicated cognitive overload, and most children maintained an appropriately balanced cognitive load over the long term, reporting significantly lower levels of training fatigue.

At the level of home-based intervention implementation, the experimental group relied on a mobile app to receive standardized home guidance, achieving a daily home-training compliance rate of 89.20%. By contrast, the control group depended solely on short weekly sessions at the institution, resulting in a home-training compliance rate of less than 15.00%, thus highlighting a clear gap in intervention continuity. Taken together, the empirical evidence confirms that the multi-source perceptual fusion intelligent intervention robot can optimize children's training conditions, enhance the long-term sustainability of interventions, and address multiple inherent limitations of traditional human-led interventions.

Discussion and implications

This study addresses four major industry pain points: a shortage of autism rehabilitation instructors in China, subjectivity in assessment, homogeneity in training, and monopoly on overseas equipment. It integrates three core theories - cognitive load theory, sensory integration theory, and natural developmental behavioral intervention - to tackle these challenges. A comprehensive, multidisciplinary, cross-disciplinary framework for intelligent intervention analysis has been established, alongside systematic and locally adapted technological solutions.

The research has led to the construction of China's first facial image dataset for children with autism, annotated with emotional and cognitive-load labels. By optimizing the AlexNet network through a dual-channel attention mechanism, the ASDface-AlexNet recognition model has been created, achieving an average emotion recognition accuracy of 89.31%. This breakthrough addresses the technical challenge of insufficient precision in general models when detecting subtle facial expressions and avoidance behaviors. A lightweight cognitive-load assessment model has been developed, integrating four types of multi-source signals: facial, ocular, head, and task-related. This model enables the classification of three levels of training states and incorporates standardized, adaptive regulation rules. This approach breaks away from the traditional training shortcomings of "one-size-fits-all" approaches at the technical level.

Furthermore, the entire system - including humanoid robot hardware and supporting collaborative mini-programs - has been integrated into a unified development platform, establishing a bidirectional, closed-loop intervention system that connects institutions and families. Over an eight-week controlled, empirical validation study, it was demonstrated that this system can significantly reduce cognitive fatigue during child training and extend the duration of rehabilitation interventions. The complete software and hardware solution possesses independent intellectual property rights and R&D potential.

Moreover, the hardware costs are substantially lower than those of comparable overseas products, making it well-suited for use by grassroots rehabilitation institutions and ordinary families with autistic children. This system offers a practical and implementable technological pathway for the digital and inclusive development of special education.

At the application and promotion level, the team will employ the Strengths, Weaknesses, Opportunities, Threats (SWOT) analysis method internally to assess the product's differentiated positioning. By examining internal strengths - including technological reserves, exclusive datasets, and interdisciplinary collaboration teams - and identifying existing weaknesses, the team will align these insights with national special education support policies. A substantial rehabilitation demand analysis will also be used to pinpoint external growth opportunities.

Additionally, by analyzing overseas competitors and rapidly iterating algorithms, the team will proactively anticipate external competitive threats. This will allow them to clearly define the core directions for robot R&D and implementation. Meanwhile, the team will utilize a variety of complementary research methods to establish a comprehensive research implementation pathway - from theoretical exploration to empirical validation.

Conclusion

On a theoretical level, this study enriches the interdisciplinary research framework at the intersection of cognitive psychology, special education, and computer vision. Most existing related studies have either focused solely on robot hardware development, single emotion recognition algorithms, or localized training modules. They fail to establish a complete logical chain encompassing "multi-source signal acquisition - quantification of cognitive states - personalized intervention output". In this study, cognitive load is taken as a mediating variable, three core rehabilitation theories are integrated with human-computer interaction technologies, and thereby the boundaries of theoretical application in intelligent rehabilitation scenarios are expanded. A specialized dataset was also constructed to fill the data gap in the niche field of affective computing within China. Meanwhile, the enhanced performance of the dual-channel attention mechanism was validated in the context of micro-expression recognition among special populations. This provides new technical insights for subsequent lightweight affective computing research.

On the practical level, the algorithmic architecture, hardware system, and service model designed in this study are not restricted by geographical boundaries. They possess universal value for promotion across special education institutions at all levels nationwide. The research findings can comprehensively alleviate the multiple real-world challenges facing the autism rehabilitation industry.

The robot can take on repetitive sensory integration and basic social skills training tasks, reducing the mechanical workload of frontline rehabilitation therapists. Its lightweight hardware is easy to deploy, making it accessible even to regions with limited rehabilitation resources, such as counties and townships. This enhances the accessibility of high-quality intervention services. The multi-source sensing system enables objective, quantitative monitoring throughout the entire training process. It automatically generates child development curves covering six key dimensions - including joint attention and fine motor skills. This addresses at the root the issue of homogenization in intervention plans. A companion mini-program provides parents with standardized home-based rehabilitation guidance, breaking away from the single-model reliance on institutional training and extending the duration of intervention.

The entire software and hardware solution is independently developed, with overall costs significantly lower than those of comparable foreign products, thus breaking the dual monopoly of imported robotic technology and pricing. This offers domestic rehabilitation institutions and ordinary families of children with autism a localized, intelligent rehabilitation solution.

All data collection and empirical experiments for this study were conducted exclusively at local rehabilitation institutions in Jinhua, China. The subject sample had limited coverage in terms of both geographic region and age stratification. Therefore, it is necessary to expand the sample size to include multiple cities and individuals with varying degrees of symptom severity in order to further validate the model's generalization performance.

Currently, the multi-source sensing system covers only four categories of behavioral indicators - facial expressions, eye movements, head gestures, and task responses - and has not yet integrated signals related to full-body limb postures or speech prosody.

Thus, there remains considerable room for expanding the dimensions of perception. Moreover, the empirical intervention period was limited to just eight weeks, lacking long-term tracking data spanning more than six months. This makes it impossible to fully observe the long-term improvement effects of robot-assisted intervention on the social and cognitive abilities of the patients. The current intervention adjustment rules are based on pre-set hierarchical logic and do not incorporate reinforcement learning to enable the device to autonomously optimize intervention strategies. There is still room for further improvement in the system's ability to make intelligent, data-driven decisions.

Subsequent research will collaborate with special education institutions and children's rehabilitation centers across multiple regions nationwide to expand the sample size, refine the stratification of the dataset by age and severity of conditions. It will also conduct multi-center, long-term, controlled trials spanning multiple regions to validate the system's adaptability and scalability for nationwide implementation.

Therefore, in the future, expanding the multimodal sensing system could be considered. This could be done by adding modules for whole-body posture capture and speech emotion recognition, as well as integrating simple physiological sensors such as heart rate monitors and skin conductance devices. Thereby, a comprehensive integrated assessment system combining physiological and behavioral data would be established. Furthermore, deep reinforcement learning algorithms might be incorporated. These algorithms could leverage the vast amount of intervention data accumulated over time. This would enable the device to autonomously learn and adapt to the optimal intervention pace and task combinations for different children. Thus, it would move away from fixed, pre-set adjustment rules.

Additionally, lightweight, iteratively developed robotic hardware could be adopted to create miniaturized home-based devices suitable for diverse settings - including ordinary home environments, inclusive kindergarten classrooms, and grassroots special education classrooms. It is also crucial to enhance the application's features, including online expert consultations and long-term rehabilitation record storage, and to establish a remote rehabilitation guidance service system.

Finally, it is essential to coordinate closely with special education equipment manufacturers and rehabilitation service platforms under the China Disabled Persons' Federation. This coordination should facilitate the practical application and commercialization of these innovations. It should also complete standardized mass production and debugging of the robots, and jointly carry out large-scale pilot programs with special education institutions across various regions. Moreover, standardized smart intervention operation manuals should be developed to support the high-quality, inclusive development of special education in China.

Funding

This work was supported by National Undergraduate Training Program on Innovation and Entrepreneurship (Grant No: 202610345074).

Acknowledgements

The authors would like to show sincere thanks to those technicians who have contributed to this research.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Ali, S., Mehmood, F., Dancey, D., Ayaz, Y., Khan, M. J., Naseer, N., Nawaz, R. (2019) An adaptive multi-robot therapy for improving joint attention and imitation of ASD children. *IEEE Access*, 7, 81808-81825.

[2] Vagnetti, R., Di Nuovo, A., Mazza, M., Valenti, M. (2026) Social robots: a promising tool to support people with autism. A systematic review of recent research and critical analysis from the clinical perspective. *Review Journal of Autism and Developmental Disorders*, 13(1), 202-226.

[3] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchwitz, A., Meneguzzi, F. (2018) Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17, 16-23.

[4] Kouroupa, A., Laws, KR., Irvine, K., Mengoni, SE., Baird, A., Sharma, S. (2022) The use of social robots with children and young people on the autism spectrum: a systematic review and meta-analysis. *PLoS One*, 17(6), e0269800.

[5] Salimi, Z., Jenabi, E., Bashirian, S. (2021) Are social robots ready yet to be used in care and therapy of autism spectrum disorder: a systematic review of randomized controlled trials. *Neuroscience & Biobehavioral Reviews*, 129, 1-16.

[6] Costley, J., Gorbunova, A., Courtney, M., Chen, O., Lange, C. (2024) Problem-solving support and instructional sequence: impact on cognitive load and student performance: J. Costley. *European Journal of Psychology of Education*, 39(3), 1817-1840.

[7] Leong, H. M., Carter, M., Stephenson, J. R. (2015) Meta-analysis of research on sensory integration therapy for individuals with developmental and learning disabilities. *Journal of Developmental and Physical Disabilities*, 27(2), 183-206.

[8] Schreibman, L., Dawson, G., Stahmer, A. C., Landa, R., Rogers, S. J., McGee, G. G., Halladay, A. (2015) Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 45(8), 2411-2428.

[9] Waddington, H., Reynolds, J. E., Macaskill, E., Curtis, S., Taylor, L. J., Whitehouse, A. J. (2021) The effects of JASPER intervention for children with autism spectrum disorder: a systematic review. *Autism*, 25(8), 2370-2385.

[10] Zhao, Z., Tang, H., Zhang, X., Qu, X., Hu, X., Lu, J. (2021) Classification of children with autism and typical development using eye-tracking data from face-to-face conversations: machine learning model development and performance evaluation. *Journal of Medical Internet Research*, 23(8), e29328.

[11] Van der Donck, S., Vettori, S., Dzhelyova, M., Mahdi, S. S., Claes, P., Steyaert, J., Boets, B. (2021) Investigating automatic emotion processing in boys with autism via eye tracking and facial mimicry recordings. *Autism Research*, 14(7), 1404-1420.

[12] Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Michie, S. (2016) Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany))*, 78(3), 175-188.

[13] Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., Yau, C. (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*, 2(10), e537-e548.