Assessing the Social Skills of Children with Autism Spectrum Disorder via Language-Image Pre-training Models

Wenxing Liu[1], Ming Cheng[1], Yueran Pan[1], Lynn Yuan[3], Suxiu Hu[3], Ming Li[1,2(✉)], and Songtian Zeng[3(✉)]

[1] School of Computer Science, Wuhan University, Wuhan, 430072, China
[2] Data Science Research Center, Duke Kunshan University, kunshan, 215316, China
ming.li369@duke.edu
[3] Shenzhen Fumi Health Technology Ltd., Co.,Shenzhen, 518000,china
songtian.zeng@umb.edu

Abstract. Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that has gained global attention due to its prevalence. Clinical assessment measures rely heavily on manual scoring conducted by specialized physicians. However, this approach exhibits subjectivity and challenges in regions without sufficient medical resources. This study presents paradigms designed to automatically evaluate various aspects of social skills in children with ASD, utilizing our multiview and multimodal behavior system. Moreover, we propose a new pipeline to predict autism-related social skill scores using the language-image pre-training model. Our multimodal behavioral database comprises 12 subjects (511 videos with labeled social skill scores). Finally, we achieve 81.46% accuracy on the paradigm success prediction task and 69.23% accuracy on the social skill ability scoring task. The results demonstrate that the language-image pre-training model can effectively introduce domain knowledge into video assessment tasks.

Keywords: Autism Spectrum Disorder · Behavior · Language-Image Pre-training Models · Social Skills Assessment.

1 Introduction

Autism Spectrum Disorder (ASD) is a prevalent neurodevelopmental disorder that commonly emerges during early childhood and significantly impacts social communication throughout an individual's lifespan [1]. The global estimate indicates a population of at least 78 million individuals affected by ASD [2].

The core characteristic of autism is a lack of social engagement and participation, along with restricted and repetitive behaviors [3]. Presently, ASD diagnosis

This research is funded in part by the Science and Technology Program of Guangzhou City (202007030011), DKU Synear and Wang-Cai Seed Grant and Shenzhen Fumi Health Technology Ltd.

primarily relies on behavior assessments, such as ADOS-G (Autism Diagnostic Observation Schedule-Generic) [4], and its revision ADOS2 [5]. Doctors typically complete a series of pre-designed social games called paradigms with children. Then, they observe the child's performance through recorded video, including social communication skills, attention span, and body movements, for 10 to 20 hours per assessment case [6]. The diagnosis is based on specialized scales and heavily relies on the subjective judgment and clinical experience of the physician [7].

Recently, the rapid development of artificial intelligence [8] and sensing technologies [9] show the potential to improve ASD assessment further. There are mainly two types of AI-based automatic assessment approaches. The first is rule-based methods. This type analyzes behavioral, social reflections, head behaviors, and eye patterns in a semi-structured paradigm and manually designs a scoring rubric from the clinical point of view to assess the child's behavioral abilities [10–12]. In this way, the paradigm design and accuracy of pattern recognition modules significantly influence the assessment performance. The second is the raw-video-based method. These methods directly use deep learning to process visual features and directly predict the labels [13, 14], suffering from the availability of training data and lack of interpretability.

This study introduces a novel methodology that utilizes language-image pre-training models to assess the social skills of ASD through paradigm videos. First, we adopt a dedicated data collection system, ensuring the presentation of standardized audiovisual stimuli and the recording of multimodal behavioral data [12]. Then, we design a series of social skill assessment paradigms inspired by the ADOS2 protocol and a behavior assessment coding rubric for each paradigm. Furthermore, we use text prompts, designed by autism domain knowledge, to guide a language-image model [15] to extract video features highly correlated with those prompts. Finally, machine learning algorithms are employed to predict the social skill ability scores of children with ASD. To our knowledge, this work is one of the first methods using the language-image pre-training model to assess the social skills of children with ASD. The contributions can be summarized as follows:

- To automatically assess the social skill of children with autism, we design and collect data for nine social skill assessment paradigms covering three areas: language, cognition, and attention.
- The proposed language-image method performs better in the social skill ability score prediction task than the rule-based and raw-video-based methods.

2 Related Works

2.1 Behavior Signal Processing System

We use a standardized platform that includes the stimulation, collection, analysis, modeling, and interpretation of human behavioral data for computer-aided

ASD diagnosis [12]. We utilize the proposed assessment environment and audio-visual analysis algorithms for rule-based approaches. Moreover, we design nine new paradigms and the associated rubric logics targeting children's different social skills.

2.2 Language-Image Pre-training Models

In recent years, pre-trained multimodal models [16–18] have achieved impressive performance on many downstream tasks, such as Image-Text Retrieval [19], Image Captioning [20], Visual Question Answering [21], etc. The contrastive language-image per-training (CLIP) [16] uses contrast learning to build connections between images and text. The vision-and-language transformer (VILT) [17] blends visual and textual inputs with a single transformer structure. To address the misalignment between visual and textual features in the semantic space, the align before fuse (ALBEF) [18] introduces a contrastive loss to align the image and text representations before fusing. However, these aforementioned models require substantial training resources, making fine-tuning them on downstream small-scale datasets challenging. In autism, the limited availability of medical data poses difficulties in supporting fine-tuning large-scale models. Our proposed method is inspired by Bootstrapping Language-Image Pre-training (BLIP) [22] and BLIP2 [15]. We utilize the lightweight Q-former to bridge the visual and language models for retrieving text-guided image features, facilitating the use of domain knowledge.

3 Methodology

3.1 Paradigm Design

Paradigms comprise a meticulously referenced series of interactive games that reflect children's social skills in language, cognition, and attention. In this study, we design a set of nine paradigms, as shown in Fig. 1. When experiments start, participants are encouraged to engage in a warm-up phase of unstructured play, allowing the children to relax and become familiar with the therapist. To reassure parents, they are invited to observe their child's performance through videos outside the assessment room. Following the warm-up phase, the child and the therapist assume seated positions on opposite sides of the table, engaging in face-to-face interaction.

Imitative Saying This paradigm assesses children's capacity to replicate the pronunciation of two-word objects[5]. The procedure is depicted in Fig. 1 a). (1) The therapist employs a slide controller to display images. A pre-selected picture, such as chopsticks, chili, mango, eggplant, or zebra, is projected onto the wall in front of the child. (2) The therapist articulates the name of the picture distinctly and audibly, only once. (3) If the child reproduces the picture's name within 3 seconds, it is considered a correct response, and the counter is added by

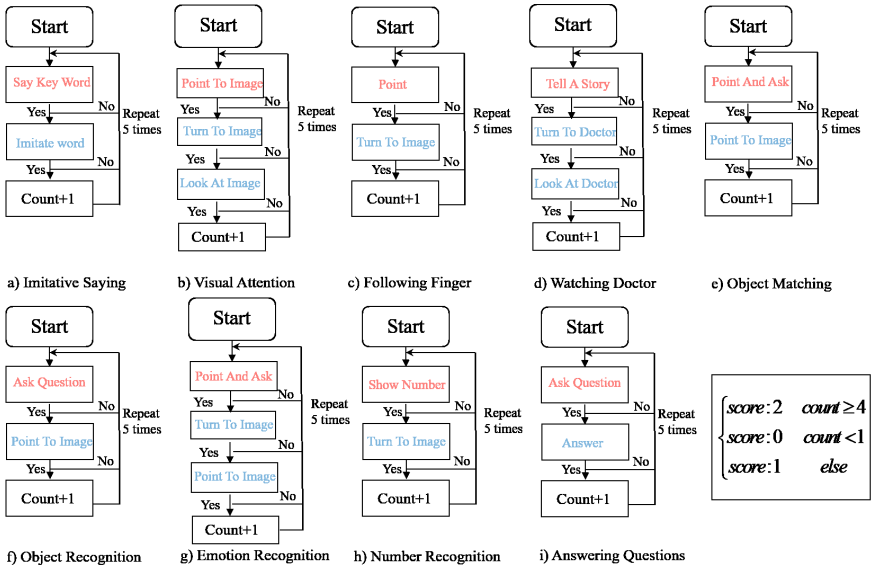

Fig. 1. The procedure of the social assessment paradigms: **Imitative Saying(IS)**, **Visual Attention(VA)**, **Following Finger(FF)**, **Watching Doctor(WD)**, **Object Matching(OM)**, **Object Recognition(OR)**, **Emotion Recognition(ER)**, **Number Recognition(NR)**, **Answering Questions(AQ)**. The words in red denote the doctor's actions, while the words in blue indicate the child's actions. Fundamental behavioral operations encompass looking, speaking, pointing, and so forth. The social assessment computer rules to determine the social skill ability score are enclosed within the box located in the lower right corner.

1. Conversely, if the child fails to respond, utters a different word, or articulates unclearly, it is deemed an incorrect behavior, and do not add to the counter (4) Repeat steps (1) - (3) five times for different pictures independently, and get this paradigm score according to assessment rubric in Fig. 1. In the rule-based automatic method, If the count is greater than or equal to 4, the computer determines the social skill ability score as 2. If the count is less than 1, the score is set as 0. In other cases, the score is considered 1.

Visual Attention This paradigm assesses the child's focus and engagement with visual stimuli, including pictures, objects, or educational material[5]. The procedure is depicted in Fig. 1 b). (1) The therapist puts the hands on the table during the initial phase, and the target image is shown on the walls. (2) The therapist raises a hand to indicate the target picture without verbal cues and then lowers the hand. (3) The therapist observes the child's reaction within 5 seconds. If the child shifts their attention to the picture and maintains looking for more than five seconds, the test is considered successful, and the counter is incremented by 1. On the contrary, incorrect behaviors encompass failure

to redirect attention towards the picture or inability to maintain continuous observation. (4) Each of the nine paradigms is repeated 5 times.

Following Finger This paradigm assesses the child's responsiveness in tracking the movement of another person's finger[5]. The procedure is depicted in Fig. 1 c). It is noteworthy that the overall process aligns with the Visual Attention paradigm. However, two notable distinctions exist: (1) The orientation of the therapist's finger does not require prior determination, thereby granting greater flexibility to the therapist. 2) There is no imposed time limit on the child's gaze toward the target. Incorrect behaviors include a lack of response, exceeding the time limit, or diverting attention toward other directions.

Watching Doctor This paradigm assesses the child's capacity to maintain continuous eye contact with the therapist during an instructional scenario[5]. The procedure is depicted in Fig. 1 d). (1) The therapist unveils a picture book and puts it on the chest. (2) The therapist delivers a story within the picture book at a natural pace, lasting approximately 10 seconds. (3) If the child redirects the gaze toward the therapist within 3 seconds, the therapist observes the child's response. A correct response is for the child to maintain the gaze on the therapist. Conversely, a lack of response, short visual engagement, or focus on alternative stimuli constitute incorrect responses.

Object Matching This paradigm is utilized to assess the child's capability for accurately matching pictures of objects [5]. The procedure is depicted in Fig. 1 e). (1) The therapist puts the hands on the table and employs a slide controller to show a picture. The target item is presented on the wall in front of the child. Two option pictures are presented on each side of the child wall. (2) The therapist raises a hand, indicates the wall containing the target, and then poses the question, "Please observe this picture and indicate which image corresponds to it." (3) If the child turns or points to a wall housing the correct object within 3 seconds, the behavior is considered successful, and the counter is added by 1. On the contrary, the incorrect behavior is that the child does not respond or points to the wrong picture.

Object Recognition This paradigm assesses children's proficiency in identifying familiar objects[5]. The procedure is depicted in Fig. 1 f). This procedure aligns with Object Matching, with the distinction lying in the second step: The therapist raises a hand and indicates the wall containing the target while posing the question, "Which one corresponds to the corn (grapes, spoons, shoes, puppies)?".

Emotion Recognition This paradigm assesses children's proficiency in recognizing the four fundamental emotions: happiness, sadness, anger, and fear.

The procedure is shown in Fig. 1 g)[5]. It is evident that the overall process closely resembles object matching. The general framework, involving presenting objects, posing questions, and awaiting responses, remains consistent. However, the content of the pictures and questions differs from that of object matching. The protocol is specifically designed to evaluate the child's emotional cognition, prompting the therapist's questions such as "Which one is happy?", "Which one is sad?", "Which one is angry?", "Which one is scared?".

Number Recognition This paradigm assesses children's aptitude in recognizing single-digit numbers[5]. The procedure is depicted in Fig. 1 h). The overall process for the paradigm is still consistent with Emotion Recognition. The overall paradigm remains consistent with Emotion Recognition. A brief description of the procedure is as follows: (1) The therapist sequentially presents a number of pictures on the walls facing the child. (2) Without any prompting, the child's successful performance is determined if he or she correctly articulates the number within 3 seconds. Conversely, failure to respond or providing irrelevant answers are considered incorrect behaviors.

Answering Questions This paradigm assesses the child's ability to provide verbal or physical responses to questions[5]. The procedure is depicted in Fig. 1 i). (1) The therapist asks the child, "Do you want to drink water?". As the rounds progress, the questions are modified accordingly (e.g., "Do you want to read a book?", "Do you want to go to the bathroom? "). (2) If the child can answer questions using verbal or body language, such as saying "Yes/No," nodding, or shaking their head, it is considered as successful. On the contrary, the incorrect reaction is responding or saying irrelevant words.

In summary, nine paradigms are designed to assess the language, cognitive, and attention abilities of children with ASD. Language paradigms include **Imitative Saying** and **Answering Questions**. Cognitive paradigms include **Object Matching**, **Object Recognition**, **Emotion Recognition**, and **Number Recognition**. Attention paradigms consist of **Visual Attention**, **Finger Tracking**, and **Teacher Observation**. Consistency has been maintained in the design of count and score ranges, with higher scores indicating stronger abilities that align more closely with typical development. Our proposed behavior signal processing system captures the ASD paradigm videos, better reflecting the social skills characteristics of children with ASD. This approach proves more relevant when extracting video features compared to using raw video data.

3.2 Language-Image Based Method

As illustrated in Fig. 2, the fusion feature extraction framework consists of three phases. (1) We first extract video frame features through the image encoder VIT-22B [23], which stands as an excellent visual model in terms of parameter count and similarity to resembles human visual perception (relying less on texture and more on the shape). (2) The Querying Transformer (Q-former) [15], an

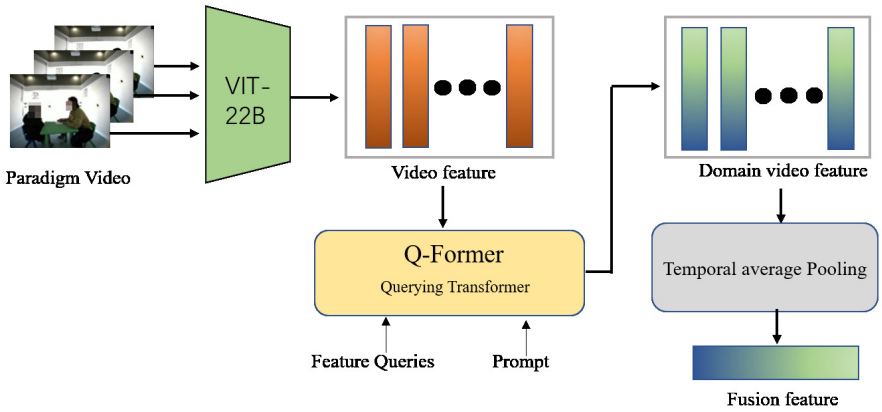

Fig. 2. Overview of language-image pre-training model base fusion feature extraction framework.

integral component in BLIP2, generates query features that interact with the text features of the prompt through the attention layer and query video features through the cross-attention layer. To summarize, Q-former allows for querying video features that exhibit semantic relevance to the text prompt, referred to as domain video features. (3) we employ a simple temporal fusion method for domain video features. Temporal Average Pooling (TAP) averages the features across the temporal channels. Hence, we obtain fused features that potentially represent the semantics of the entire video.

In experiments, we directly employ the pre-trained VIT-22B and Q-former models for inference. Each paradigm video is encoded as a 768-dimension fused feature. Subsequently, we employ the support vector machine (SVM) with the linear kernel as a classifier to predict the social skill paradigms count and ability score in a leave-one-subject-out manner. An essential aspect of this method lies in the prompt design: The American Psychiatric Association's Diagnostic and Statistical Manual Fifth Edition(DSM-5) provides standardized criteria to help diagnose ASD [24]. We design text prompts regarding the DSM-5 to integrate domain knowledge with video features, as shown in Table 1.

4 Experimental Results

4.1 Database

We recruit 12 participants, aged between 3 and 6 years, consisting of seven individuals diagnosed with autism and five typically developing children. Informed consent was obtained from the parents of all children prior to commencing the formal assessment. We created a dataset containing 9 paradigms for 511 paradigm videos (58 for IS, 60 for VA, 36 for FF, 60 for WD, 60 for OM, 60 for OR, 60 for ER, 57 for NR, and 60 for AQ). Each paradigm video has 8 views, and our method uses only one view when extracting fusion features.

Table 1. Details of DSM-5 Prompts

Index	Prompt Context
1	The child of Deficits in social-emotional reciprocity
2	The child of Deficits in nonverbal communicative behaviors used for social interaction.
3	The child of Deficits in developing, maintaining, and understanding relationships.
4	The child of Stereotyped or repetitive motor movement.
5	The child of Insistence on sameness, inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior.
6	The child of Highly restricted, fixated interests that are abnormal in intensity or focus.
7	The child of hyper- or hyporeactivity to sensory input or unusual interest in sensory aspects of the environment.

To minimize subjectivity, we engaged three professional therapists as evaluators to score the collected video database during the manual labeling process. Each evaluator underwent a training session to familiarize themselves with the rubrics for coding paradigm performance. Additionally, they independently reviewed each video recording in the database, and the majority vote from the three evaluators was used as the ground truth for paradigm successes and ability scores. Our use of data is approved by our Institutional Review Board(IRB).

4.2 Results

We adopt two metrics to evaluate our automatic assessment methods, namely Paradigm Success Accuracy (P-Acc) and Ability Score Accuracy (A-Acc). P-Acc, representing the performance of a two-class classification task, signifies the accuracy rate of an individual single-round paradigm success assessment. A-Acc, representing the performance of a three-class classification task, indicates the accuracy rate of predicting the child's social skill ability scores.

We conduct comparative experiments on the proposed dataset with rule-based and raw video-based methods. The rule-based method uses the Multiview and Multimodal Behavior Transcription (MMBT) system to recognize fundamental human behaviors from recorded data [12]. The raw video-based method extracts features directly from raw videos, and also selects VIT-22B as the feature extractor and SVM as the classifier to ensure the fairness of the experiment.

As the number of video recordings in each paradigm is small, leave-one-subject-out cross validation [25] is adopted to obtain the accuracy estimation for two classification tasks. In our leave-one-subject-out cross-validation, we leave one child's data as the test data and other children's data as the train data for each paradigm. The data for each paradigm is independent, and the training and testing process does not use data from other paradigms.

Table 2. The performance of our proposed pre-training model based method compared to the rule-based and raw-video-based methods. P-Acc and A-Acc denote the Paradigm Success Accuracy and Ability Score Accuracy, respectively.

Paradigm	Rule-Based		Raw-Video-Based		Ours	
	P-Acc	A-Acc	P-Acc	A-Acc	P-Acc	A-Acc
IS	88.33%	75.00%	51.72%	29.31%	84.48%	65.51%
VA	85.56%	75.00%	41.66%	53.33%	70.00%	78.33%
FF	86.25%	66.67%	83.33%	22.22%	91.66%	81.67%
WD	80.56%	75.00%	72.22%	38.88%	88.88%	66.67%
OM	80.00%	33.33%	43.33%	21.67%	70.00%	48.33%
OR	81.67%	66.67%	38.33%	31.67%	71.67%	66.66%
ER	80.00%	75.00%	61.67%	30.00%	80.00%	58.33%
NR	74.03%	33.33%	82.45%	63.16%	96.49%	84.21%
AQ	83.33%	66.67%	70.00%	40.00%	80.00%	73.33%
average	82.20%	62.96%	60.52%	36.69%	81.46%	69.23%

Table 2 compares the results of our proposed method and two baselines from both P-Acc and A-Acc perspectives. From the P-Acc viewpoint, our method achieves an average accuracy of 81.46% across the 9 paradigms, which approaches the 82.20% accuracy achieved by the rule-based method. Notably, the rule-based method utilizes multiple additional high-precision sensors (RGB-D cameras) and high-accuracy behavior signal processing re-trained models [12] (human detection, gaze, head pose, hand gesture, speech recognizer, etc.) Our proposed method achieves a comparable level of accuracy utilizing the RGB video data solely, without using the paradigm design rubrics. In addition, our proposed method improves accuracy by 20.94% compared to the raw-video-based method. Regarding the A-Acc perspective, the rule-based method only attains a score of 62.96% due to the constraints imposed by manually defined ability scoring rules, despite having a higher paradigm accuracy rate. Our method achieves the highest average ability accuracy of 69.23% across the 9 paradigms, outperforming the two baselines by 6.27% and 32.54%, respectively.

The highest P-Acc for our proposed method is 96.49% on the Number Recognition (NR) paradigm, and the highest A-Acc for our proposed method is 81.67% on the Visual Attention (VA) paradigm. The reason for this result may be related to the design of the prompt, where our approach introduces attention-related prompts, e.g. (6) in Table 1. The ability to focus is a common feature of both paradigms. In the NR paradigm, children need to attend to the positions of numbers, whereas, in the VA paradigm, their attention is directed toward the positions of pictures.

4.3 Discussion

In this paper, we propose a language-image pre-training model based behavior assessment method, which incorporates domain knowledge inspired prompts in the feature extraction process. The method assesses the similarity between

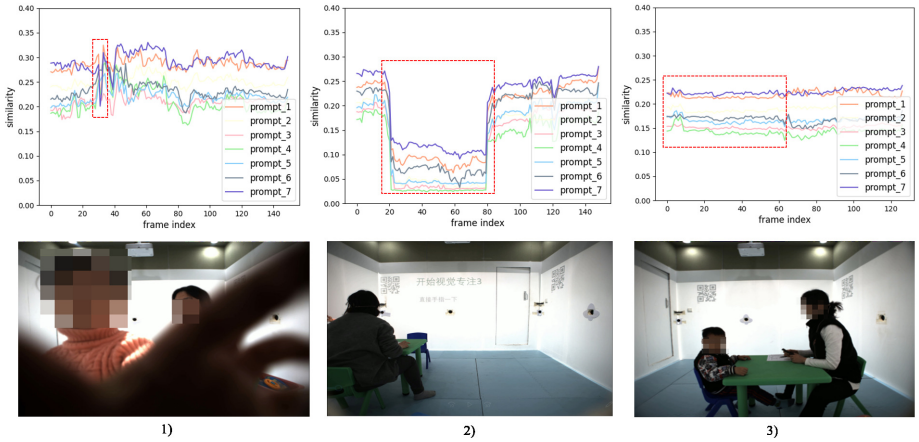

Fig. 3. Three instances of prompt similarity curves corresponding to real scenes. The Similarity meaning is the cosine similarity between image and prompt features. The images in the first row are the prompt similarity curves crossing time, and the second row is one real scene image in the red box.

prompt features and image features for each video frame, allowing for the retrieval of features that capture prompt semantics. Consequently, two essential questions arise: 1) Does the change in similarity correspond to events in real-life scenarios? 2) Can the content of the prompt partially explain the paradigm?

To address the first question, Fig. 3 provides three examples. In scenario 1), the similarity curve of the red box exhibits a sharp decline, indicating a realistic situation where children suddenly cover the camera with their hands. In scenario 2), the similarity curve of the red box reaches a significantly low level for a period, corresponding to instances where the child moves away from the camera scene. The smooth variation observed in scenario 3) aligns with children's sustained attention to pictures in the VA paradigm. The child's disappearance in the actual scenario leads to a decrease in correlation, and the continued focus of the child leads to a stable value in correlation. This illustrates that the similarity between prompts and pictures reflects the true variability of the scene. Furthermore, children with ASD generally exhibit higher similarity average value compared to normal children. These observations validate the effectiveness of the Language-Image method in extracting domain-specific features since our prompts are common symptoms of ASD.

To address the second question, we analyze the correlation rankings of different prompts for each paradigm, as presented in Table 3. The content of the prompt contributes to explaining the paradigm to some extent. For instance, the child's attention is a shared observation across the VA, FF, WD, and OR paradigms. Among all the prompts, prompt 6 stands out as the most representative in terms of autistic attention. Consequently, prompt 6 exhibits the highest correlation, as anticipated.

Table 3. The correlation ranking of different Prompts for each paradigm. The Correlation ranking is organized in descending order. P-order means prompt ordering in paradigm assessment, and A-order means prompt ordering in ability score assessment.

Paradigm	IS	VA	FF	WD	OM	OR	ER	NR	AQ
P-Order	5164237	2543671	6124735	1657324	3672145	6574213	5732614	2315467	4726153
A-Order	1476325	6273145	6123457	4671235	2845376	6521347	3746125	6532174	7263451

5 Conclusion

This work designs nine social skills assessment paradigms covering language, cognition, and attention domains. Additionally, we introduced a novel language-image pre-training model based approach to predict the social skill assessment labels. Our method achieves a paradigm success accuracy of 81.46% and an ability score accuracy of 69.23% on our dataset. Notably, our proposed method outperforms the rule-based and raw-video-based methods by incorporating domain knowledge into the text prompts while relying solely on the RGB video data. It has great potential to avoid manually designing the social skill paradigm assessment rubrics and complex systems built upon many behavior signal processing modules.

References

1. Lord, C., Charman, T., Havdahl, A., Carbone, P., Anagnostou, E., Boyd, B., McCauley, J. B.: The Lancet Commission on the future of care and clinical research in autism. The Lancet. 399(10321): 271-334(2022)
2. Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., Dowling, N. F.: Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network. MMWR Surveillance Summaries. 67(6): 1(2018)
3. Harris, J.: Leo Kanner and autism: a 75-year perspective. International review of psychiatry. 30(1), 3-17(2018)
4. Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Rutter, M.: The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of autism and developmental disorders. 30, 205-223(2000)
5. Gotham, K., Risi, S., Pickles, A., Lord, C.: The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. Journal of autism and developmental disorders. 37, 613-627(2007)
6. Falkmer, T., Anderson, K., Falkmer, M., Horlin, C.: Diagnostic procedures in autism spectrum disorders: a systematic literature review. European child & adolescent psychiatry. 22, 329-340(2013)
7. Taylor, L. J., Eapen, V., Maybery, M., Midford, S., Paynter, J., Quarmby, L., Whitehouse, A. J.: Brief report: An exploratory study of the diagnostic reliability for autism spectrum disorder. Journal of autism and developmental disorders. 47, 1551-1558(2017)

8. de Belen, R. A. J., Bednarz, T., Sowmya, A., Del Favero, D.: Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. Translational psychiatry. 10(1), 333(2020)
9. Winoto, P., Chen, C. G., Tang, T. Y.: The development of a Kinect-based online socio-meter for users with social and communication skill impairments: a computational sensing approach. In Proc. ICKEA, pp. 139-143(2016)
10. Bovery, M., Dawson, G., Hashemi, J., Sapiro, G.: A scalable off-the-shelf framework for measuring patterns of attention in young children and its application in autism spectrum disorder. IEEE transactions on affective computing. 12(3), 722-731(2019)
11. Wang, Z., Liu, J., He, K., Xu, Q., Xu, X., Liu, H.: Screening early children with autism spectrum disorder via response-to-name protocol. IEEE Transactions on Industrial Informatics, 17(1), 587-595(2019)
12. Cheng, M., Zhang, Y., Xie, Y., Pan, Y., Li, X., Liu, W., Li, M.: Computer-Aided Autism Spectrum Disorder Diagnosis With Behavior Signal Processing. IEEE Transactions on Affective Computing. (2023)
13. Li, J., Zhong, Y., Han, J., Ouyang, G., Li, X., Liu, H.: Classifying ASD children with LSTM based on raw videos. Neurocomputing, 390, 226-238(2020)
14. Negin, F., Ozyer, B., Agahian, S., Kacdioglu, S., Ozyer, G. T.: Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. Neurocomputing, 446, 145-155(2021)
15. Li J, Li D, Savarese S, et al.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597(2023)
16. Radford A, Kim J W, Hallacy C, et al.: Learning transferable visual models from natural language supervision. In Proc. ICML, pp. 8748-8763(2021)
17. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In Proc. ICML, pp. 5583-5594(2021).
18. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S. C. H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34, 9694-9705(2021)
19. Cao M, Li S, Li J, et al.: Image-text retrieval: A survey on recent research and development. arXiv preprint arXiv:2203.14713(2022).
20. Stefanini M, Cornia M, Baraldi L, et al.: From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence, 45(1): 539-559(2021)
21. Lin Z, Zhang D, Tao Q, et al.: Medical visual question answering: A survey. Artificial Intelligence in Medicine, 102611(2023)
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proc. ICML, pp. 12888-12900(2022)
23. Dehghani M, Djolonga J, Mustafa B, et al.: Scaling vision transformers to 22 billion parameters. arXiv preprint arXiv:2302.05442(2023)
24. American Psychiatric Association D, American Psychiatric Association.: Diagnostic and statistical manual of mental disorders. American psychiatric association. Washington, DC(2013).
25. Wong T T.: Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition. 48(9), 2839-2846(2015).