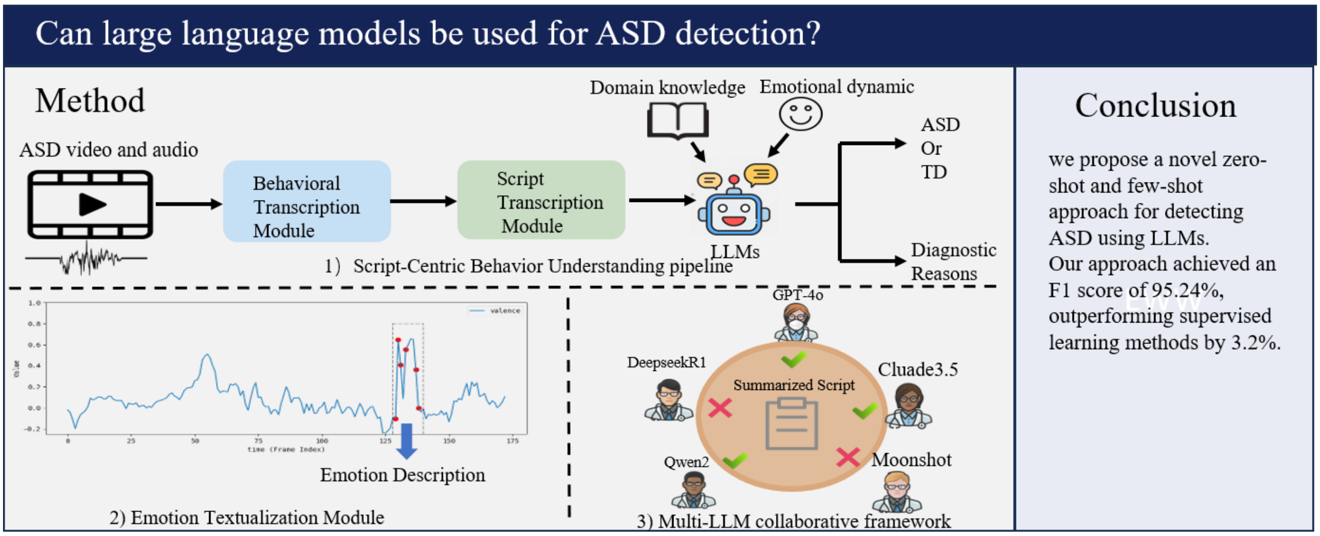
Graphical Abstract

Detecting Children with Autism Spectrum Disorder based on Script-Centric Behavior Understanding with Emotional Enhancement

Wenxing Liu, Yueran Pan, Dong Zhang, Hongzhu Deng, Xiaobing Zou, Ming Li



Highlights

Detecting Children with Autism Spectrum Disorder based on Script-Centric Behavior Understanding with Emotional Enhancement

Wenxing Liu, Yueran Pan, Dong Zhang, Hongzhu Deng, Xiaobing Zou, Ming Li

- To the best of our knowledge, our approach is the first to introduce LLMs for detecting ASD from audio-visual data, laying the foundation for the exploration of LLMs in this domain.
- In order to accurately describe the behavioral data and utilize the domain knowledge, we propose a script transcription module and a domain prompt module. They build a bridge between audio-visual data and LLMs, facilitating the development of multimodal ASD detection.
- We propose SCBU-Agents, a multi-LLM collaborative framework that improves ASD detection accuracy by enabling re-analysis, discussion, and consensus-building among diverse large language models.
- We design an emotion textualization module to add emotional dynamics that are usually ignored in the behavioral script, further enhancing the detection accuracy.

Detecting Children with Autism Spectrum Disorder based on Script-Centric Behavior Understanding with Emotional Enhancement[⋆,⋆⋆]

Wenxing Liu[a,b], Yueran Pan[a,b], Dong Zhang[c], Hongzhu Deng[d], Xiaobing Zou[d] and Ming Li[a,b,*]

[a] *the School of Computer Science, Wuhan University, Wuhan, 430072, Hubei, China*
[b] *the Digital Innovation Research Center, Duke Kunshan University, 215316, Kunshan, Jiangsu, China*
[c] *the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510006, Guangdong, China*
[d] *the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, 510630, Guangdong, China*

ARTICLE INFO

Keywords:
Autism Spectrum Disorder, Behavior Textualization, Emotion Textualization, Large Language Model

ABSTRACT

Early diagnosis of autism spectrum disorder (ASD) is critical but often hindered by clinician scarcity and subjective interpretation of social behaviors. While machine learning can assist, it typically demands large, labeled datasets and lacks interpretability. To address these limitations, we introduce a zero-shot and few-shot framework for ASD detection based on Script-Centric Behavioral Understanding (SCBU) with emotional enhancement. Our method transforms multi-view audio-visual recordings of standardized social interactions into structured, time-stamped behavioral scripts. These scripts encapsulate gaze, gestures, head pose, and speech, augmented by an emotion textualization module that captures dynamic affective changes. A domain prompt module integrates clinical knowledge, and the resulting scripts are analyzed by Large Language Models (LLMs) to infer ASD status and generate diagnostic rationales. We further propose SCBU-Agents, a multi-LLM collaboration strategy that achieves consensus through re-analysis and discussion. Evaluated on a clinical dataset of 95 children, our framework achieved an F1-score of 95.24%, outperforming both behavior signal processing baselines and human raters while preserving interpretability. This work demonstrates that textualizing behavior and leveraging LLM reasoning can provide accurate, interpretable, and privacy-preserving clinical decision support for ASD screening.

1. Introduction

Autism is a neurodevelopmental disorder characterized by impairment in communication and social interaction, restricted interests and stereotyped behaviors [43, 67]. It imposes substantial psychological and financial burdens on families and poses broader societal challenges [46].

Numerous studies show that behavioral interventions for individuals with ASD can effectively mitigate core symptoms when initiated during key developmental periods [68]. Accurate early diagnosis is essential for the subsequent intervention. However, conventional gold-standard diagnostic tools require highly experienced clinicians and are time-consuming, like the Autism Diagnostic Interview-Revised (ADI-R) [44] and the Autism Diagnostic Observation Schedule (ADOS) [45], typically rely on interviews or behavioral observations. These methods are inherently subjective due to variations in clinicians' interpretations and differences in individual expertise. Hence, the implementation of these diagnostic protocols continues to pose significant challenges, particularly in resource-constrained healthcare regions. This underscores the significant and critical need for developing an objective, accurate and interpretable early screening tool for ASD [85].

To address these problems, recent research demonstrates the effectiveness of artificial intelligence techniques for ASD detection using multimodal sensor data, e.g. functional Magnetic Resonance Imaging (fMRI) data [26, 57, 40], Electroencephalogram (EEG) data [3, 53, 90], behavioral observation audio-visual data [99, 10, 15], etc. This paper focuses on automated ASD detection through audio-visual data analytics. Conventional behavior data-based methods can be divided into two main categories, as shown in Fig. 1 (a) and (b). The first is based on modular behavioral signal processing. These methods extract the child's behavioral features (e.g., gaze patterns [102], head movements [35], body motions [31], audio prosody and text transcripts [19], facial expression and hand gestures [38], etc.) and then analyze these features by backend machine learning algorithms. The second is based on end-to-end raw video classification, which directly learns the mapping relationship between video and labels through deep learning models [60, 64, 92]. Both methods provide quantitative and objective diagnostic tools for ASD detection. However, these supervised learning methods need to be trained by a large amount of data,

[⋆] This research is funded in part by the Guangdong Science and Technology Plan (2023A1111120012), and National Natural Science Foundation of China (62171207, 62173353).

[⋆⋆] The collection and analysis of the clinical database was approved by the Third Affiliated Hospital of Sun Yat-sen University Institutional Review Board (IRB No. [2018]02-196-01 and [2020]02-118-03) and Duke Kunshan University Institutional Review Board (IRB No. 2022ML065,2023LM159, 2025LM039). Access to the audio-visual data remains restricted, but we are prepared to share and contribute towards collaborative efforts concerning the available behavior scripts data.

[*] Corresponding author
ming.li369@dukekunshan.edu.cn (M. Li)
ORCID(s): 0000-0002-6406-1983 (M. Li)

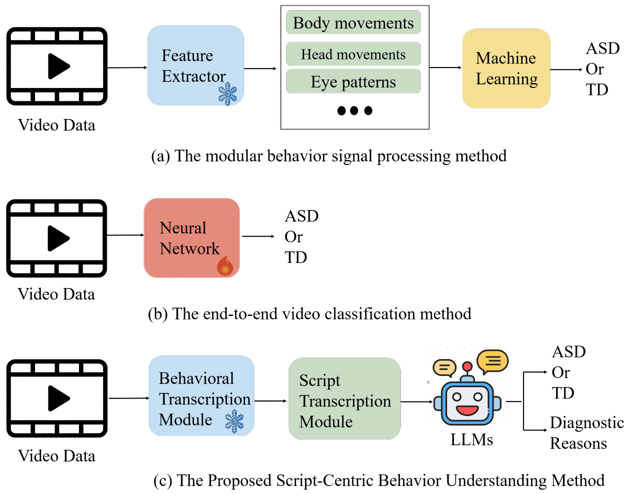


Figure 1: Pipeline comparing (a) behavioral signal processing method , (b) raw-video-based method and (c) The proposed script-centric behavior understanding (SCBU) method

especially for the end-to-end video classification approach. Hence, the scarcity of ASD data limits their accuracy in practice. Moreover, most of the aforementioned methods are limited to making binary predictions and lack detailed explanations supporting those detection outcomes.

In recent years, researchers have gradually utilized LLMs [91, 98, 22] to explore new approaches for medical diagnosis. Compared to traditional AI algorithms, LLMs exhibit two primary advantages for clinical applications [101, 82, 56]. First, their ability to leverage extensive medical knowledge enables reasoning and contextual comprehension, forming an interpretable foundation for diagnostic decision-making. This capability establishes a crucial foundation for ensuring the interpretability of diagnostic decisions. Second, their inherent prior knowledge supports zero-shot and few-shot learning, reducing dependence on clinical training data.

While LLMs show potential for medical diagnostics, three fundamental challenges persist in processing ASD-specific audio-visual behavioral signals: 1) Modality adaptation gap. Current LLMs excel at text and image processing [27, 2, 18, 94, 50, 23], but face inherent limitations in accurate and robust cross-modal alignment between audio-video inputs and textual representations [36, 97], creating fundamental barriers to biomedical behavioral signal processing. 2) Domain prompt Design. While prompt engineering significantly enhances LLM capabilities for domain-specific tasks [89], the systematic design of effective prompts and the optimal infusion of ASD-related domain knowledge still remain important research questions. 3) Emotional dynamic Modeling. Emotional dynamic is a critical biomarker for ASD detection [8, 48, 65], a key challenge lies in modeling emotional dynamic as a textual feature that enhances LLMs-based ASD detection.

Inspired by social cognitive psychology [71] and narrative psychology [49], we transform multimodal behavioral records into natural language scripts. By textualizing behaviors, the SCBU framework abstracts multimodal data into a semantic-level representation interpretable by LLMs, enabling symbolic reasoning about intention, affect, and reciprocity rather than low-level feature correlation. This approach also aligns with computational theory-of-mind models [12], where textual narratives serve as a bridge between perceptual cues and mental-state inference. Furthermore, substantial domain knowledge and prior research already exist in the textual form. Behavioral textual representations enable LLMs to more effectively leverage this rich body of literature, including books, scientific papers, and DSM-5 diagnostic criteria, etc.

Therefore, this paper introduces a novel pipeline for ASD detection using LLMs, which can determine ASD and provide explanations. First, to solve the problem of mismatch between audio-visual behavior data and the input modality of LLMs, we convert the video content (e.g. the characters' gestures, head poses, body' movements, facial expressions, speech content, gaze, etc.) into time-stamped textual descriptions. Specifically, we use a behavioral transcription module to convert the video content into human behavioral logs, and a script transcription module is designed to process these behavioral logs into natural language texts. Second, to improve the accuracy of LLM for ASD detection, we created a domain prompt module to incorporate ASD domain knowledge. Finally, we design the emotion textualization module to enhance LLMs' understanding of the emotional dynamics in ASD detection. Our main contributions can be summarized as follows:

- To the best of our knowledge, our approach is the first to introduce LLMs for detecting ASD from audio-visual data, laying the foundation for the exploration of LLMs in this domain.
- In order to accurately describe the behavioral data and utilize the domain knowledge, we propose a script transcription module and a domain prompt module. They build a bridge between audio-visual data and LLMs, facilitating the development of multimodal ASD detection.
- We propose SCBU-Agents, a multi-LLM collaborative framework that improves ASD detection accuracy by enabling re-analysis, discussion, and consensus-building among diverse large language models.
- We design an emotion textualization module to add emotional dynamics that are usually ignored in the behavioral script, further enhancing the detection accuracy.

2. Related Work

2.1. Modular Behavior Signal Processing Methods

To overcome ASD data scarcity, some approaches leverage cross-domain behavioral signal processing modules for feature extraction, followed by ASD-targeted modeling with domain knowledge. Hashemi et al. [24] developed a mobile screening system that employs cinematic stimuli to elicit quantifiable social responses (e.g., name recognition, joint attention, affective reciprocity) for automated ASD classification via behavioral pattern analysis. Negin et al. [52] implemented a Bag-of-Visual-Words (BoVW) framework to extract local descriptors from video data, employing multiple machine learning classifiers for automated ASD screening. Zhang et al. [100] implemented a 3D spatiotemporal facial analysis pipeline and a few-shot learning strategy to evaluate discriminative facial dynamics for ASD classification. Atyabi et al. [5] developed a multimodal integration framework combining behavioral biomarkers—including eye movement scan paths, temporal information and pupil velocity—to differentiate ASD and Typically Developed (TD). Cheng et al. [10] developed a computer-aided ASD detection system employing multimodal behavioral signal analysis. The system's multimodal behavioral transcription module and response parser recognizes audio-visual signals to identify child's behaviors, and a back-end machine learning model is trained based on the paradigm scores and behavioral features to provide assisted ASD detection. Nie et al. [9] formalized child-caregiver interactions through a Computational Interpersonal Communication Model (CICM) grounded in Theory of Mind (ToM), employing Markov decision processes to decode multimodal behavioral signals for early screening. While demonstrating diagnostic potential, these approaches often require considerable domain-specific expertise and customized behavioral signal processing module designs, exhibiting limited generalizability across diverse clinical scenarios.

2.2. End-to-end deep learning-based methods

Many studies employ end-to-end deep learning models to analyze raw videos directly. Tang et al. [80] propose an automated early-screening framework that analyzes infants' visual and vocal behaviors during the Still-Face Paradigm, extracting head-movement, facial-appearance, and acoustic features from video/audio recordings. Li et al. [37] pioneered an LSTM-based deep learning architecture for automated ASD detection through raw video, specifically targeting discriminative gaze pattern classification between ASD and TD. Pandian et al. [69] developed the RGBPose-SlowFast network to automatically detect stereotypical motor behaviors in children with ASD, demonstrating the viability of multi-stream neural networks for automated ASD screening. Wei et al. [88] explored a hybrid architecture integrating handcrafted feature extractors with 3D convolutional neural networks for automated detection of stereotyped motor mannerisms in ASD. Chola Raja and Kannimuthu [11] introduced a hybrid meta-heuristics model that combines Seagull and Elephant Herding Optimization for deep feature selection, achieving robust performance under supervised training. Aarthi and Kannimuthu [1] provided a comprehensive survey summarizing more than fifty machine-learning approaches and outlining current progress in ASD prediction using both classical and deep models. Raja et al. [63] further proposed a conditional GAN-based framework that enhances ASD data diversity through synthetic sample generation, demonstrating the potential of generative modeling in clinical data augmentation. Chen et al. [92] employ Longformer to establish the correlation of facial features in videos over time, aiming to learn a deep representation from dynamic facial data for ASD detection. Asha et al. [64] developed a supervised contrastive learning framework to extract cross-dataset discriminative feature representations for ASD and TD, ultimately deploying an automated diagnostic classification pipeline leveraging raw video. Khan and Katarya [28] integrates Bat and Particle Swarm Optimization with an LSTM network using adaptive feature fusion, improving training convergence and classification stability. Khan and Katarya [29] develops a multimodal transformer that jointly encodes video, audio, and textual modalities via cross-attention, enabling richer cross-modal interaction modeling. Khan and Katarya [30] combines White Shark Optimization with Bi-LSTM to better capture temporal dependencies in behavioral sequences and enhance diagnostic precision. Despite their methodological diversity, these studies share common limitations. They rely heavily on supervised training with task-specific datasets, offer limited interpretability, and rarely model the emotional or contextual dynamics essential to social interaction analysis.

2.3. LLMs Based Methods

In recent years, LLMs have demonstrated advanced natural language processing capabilities, showing particular promise in clinical diagnostics for developing interpretable decision-making systems that balance accuracy and transparency. The Med-PaLM [73] is a medical LLM developed by the Google team. Its strong performance in answering medical questions demonstrates the extensive medical knowledge embedded in LLMs. Huatuo GPT [86] is an LLM for medical consulting built on the open-source LLaMa-7B model [96]. The LLM learns structured and unstructured medical knowledge from the Chinese Medical Knowledge Atlas and can diagnose over 3,000 diseases. Through a prompt engineering approach, Medprompt [54] demonstrates that general LLMs achieve competitive performance even without domain-specific pretraining. The LLaVA-ASD [16] multimodal LLM is designed to detect social and repetitive behaviors through audio-visual cues. While multimodal LLMs can process audio-visual inputs directly, insufficient detail in behavioral descriptions generated by general visual understanding models lead to reduced performance.

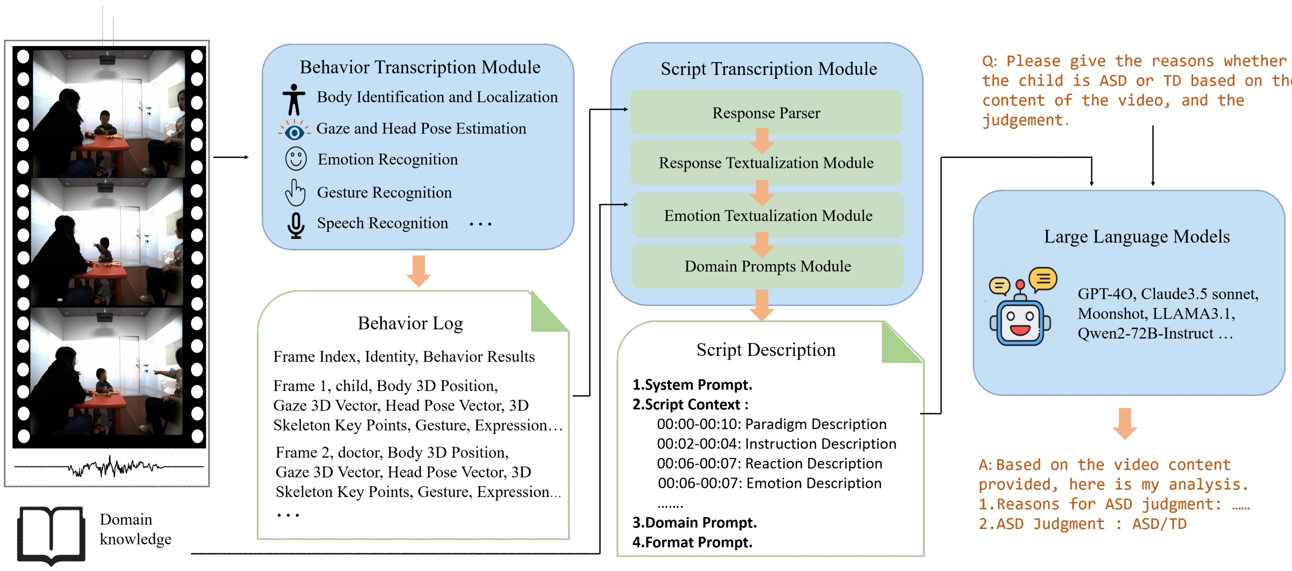


Figure 2: The overview of our proposed Script-Centeric Behavior Understanding(SCBD) framework. **Behavior Transcription Module** converts audio-video data into behavioral logs using multiple well-trained behavior signal processing models. **Scipt Transcription Module** textualizes Behavior Logs in stream and integrate domain prompt. **Large Language Models** are used to understand and anwser script content.

2.4. Emotion Based Methods

The DSM-V [67], published by the American Psychiatric Association, is the most authoritative and widely used manual for diagnosing mental disorders, including ASD. Its criteria emphasize significant deficits in the emotional domain, such as lack of facial expression. These challenges stem from impaired emotional expression and regulation mechanisms in ASD [8]. This has driven the development of computational approaches targeting emotion-related deficits in ASD through computer-assisted methodologies. Sarabadani et al. [70] proposed a method to recognize emotional states by physiological signals automatically. The study indicated that children with ASD exhibited different responses compared to children with TD when viewing images of the same emotional valence. Piana et al. [58] developed an automatic emotion recognition system to support children with ASD in learning emotion recognition and expression through whole-body movements. Prakash et al. [60] develop a framework for extracting motor behaviors, emotional states, and facial expressions from child-caregiver interactive videos. This multimodal integration of behavioral and affective data enables robust diagnostic frameworks for ASD. Rashidan et al. [66] verified that appropriate video stimuli can elicit emotional responses in children with ASD and also demonstrated significant differences in emotion regulation between children with ASD and TD.

In summary, given the limited audio-visual data in the ASD domain, using behavior signal processing modules, including emotion recognition modules, to generate high quality behavior scripts and feeding them to LLMs with ASD domain knowledge for zero-shot or few-shot learning is a very promising area to explore for ASD detection.

3. Method

Fig. 2 illustrates the overall framework of our proposed Script-Centric Behavior Understanding (SCBU) method, enabling LLMs to detect ASD from audio-visual behavior data automatically. The framework follows a sequential structure consisting of three modules: Behavioral Transcription Module, Script Transcription Module and LLMs.

The Behavioral Transcription Module recognizes basic human behaviors from audio-video data. The Script Transcription Module, newly proposed in this work, bridges the gap between behavioral logs and LLMs. It comprises four components: 1) The response parser module captures predefined response signs. 2) The response textualization module converts behavior activity into a script (text format). 3) The emotion textualization module enhances script details by adding appropriate emotional descriptions. 4) The domain prompts module combines script context, system prompt and ASD-related knowledge together to understand human behaviors better. Ultimately, we rely on pretrained LLMs to detect ASD and produce judgments by answering questions based on the script description.

3.1. Behavior Transcription Module

Following our previous work [10], the behavior transcription module can recognize the position, gesture, body movement, head pose, eye gaze, emotion, speech content of all individuals in each segment by combining audio and image models. As shown in Fig. 3, the behavioral transcription module is divided into two phases to convert recorded audio-visual data into behavioral logs.

The first stage, called the multi-person identification and localization stage, has the core objective of localizing

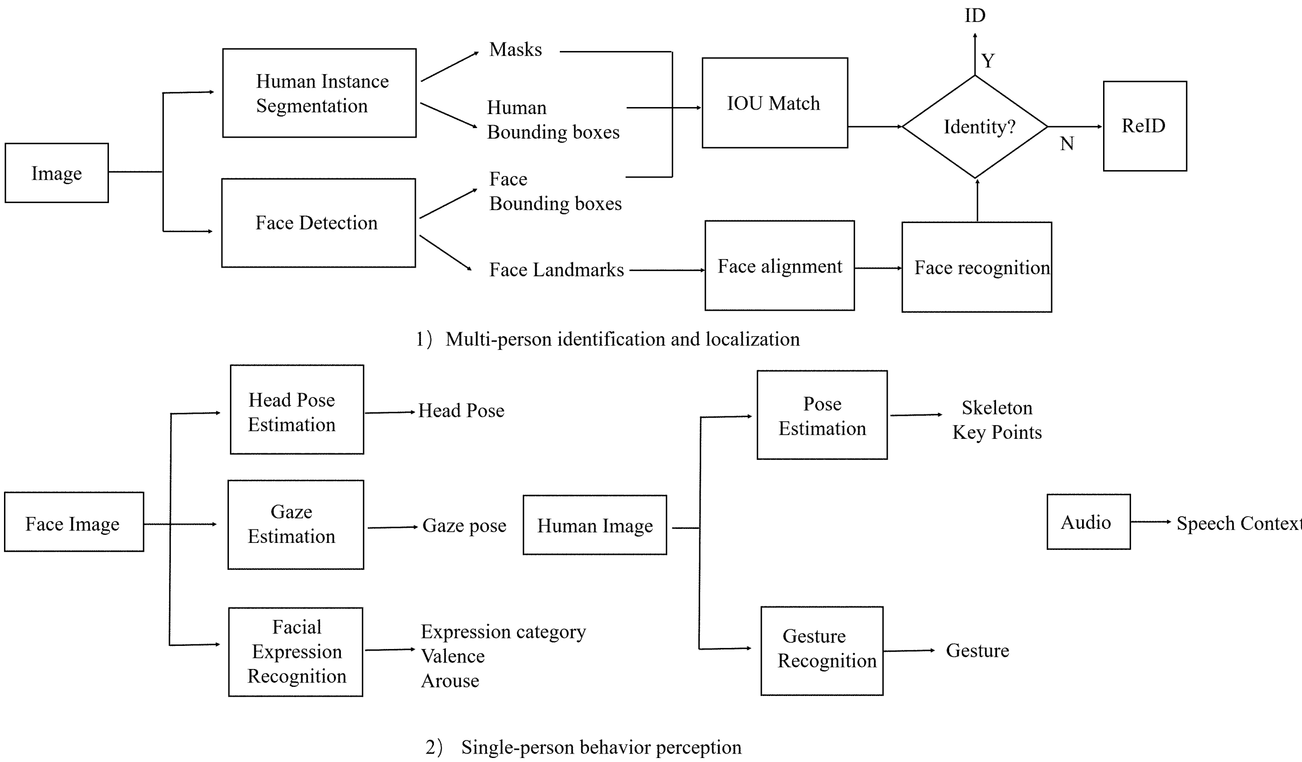


Figure 3: A two-stage pipeline in the behavioral transcription module. (1) **Multi-person identification and localization** is used to locate the location information and identify the participant in each frame. (2) **Single-person behavior perception** is used to perceive the behavioral information of each individual

and identifying different people in each frame. For each image frame, we first use an instance segmentation model (Solov2 [87]) to extract human body regions, including bounding boxes and masks. Moreover, it is further processed by a face detection model (RetinaFace [13]) to localize face bounding boxes. To associate each face bounding box with a body region, we select the match with the minimal Intersection over Union (IoU). Then, the face recognition model (ArcFace [14]) is used to distinguish the identity of characters. If face recognition is not possible due to facial occlusion, we use the person re-identification model (BOTR-ReID [47]) to recognize the identity from body features.

The second stage, called the single-person behavior perception stage, processes the behavior of the identified characters separately. The gaze and head pose estimation model (SYSUGaze[39]) are used to localize the direction of a character's head and gaze. The emotion recognition model [95] is used for three-category facial expression classification (neutral, happy, sad) and continuous valence and arousal emotion regression. The body key points model (HRNet[77]) is used to recognize hand-raising movements. The Yolov5[1] model is used to detect the hand region and train a standard ResNet-50 model [25] to recognize the 4-class gesture. The Automatic speech recognition model (Kaldi[59]) is used to recognize participants' speaking contents.

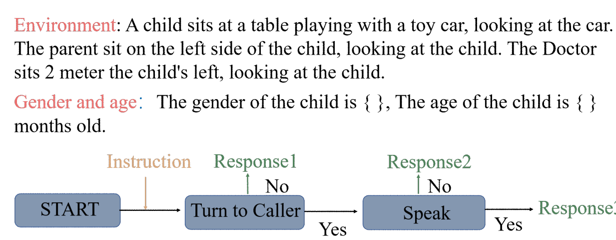


Figure 4: The textualization process of the Response to Name paradigm.

We formulate audio-visual transcription as a multi-model perception task. This is denoted as:

$$[B]_{\mathrm{i}} = \mathrm{f}^{j}_{\text{image}}\left(I_i\right) + \mathrm{f}^{\mathrm{k}}_{\text{audio}}\left(S_{\mathrm{i}}\right) \tag{1}$$

where $[B]_{\mathrm{i}}$ denotes the behavioral log of the i_{th} frame, $\mathrm{f}^{j}_{\text{image}}$ denotes the j_{th} image model, $\mathrm{f}^{\mathrm{k}}_{\text{audio}}$ denotes the k_{th} audio model, where $\left(I_i\right)$ and $\left(S_i\right)$ denote the i_{th} frame of image and audio signals, respectively.

To ensure a fair comparison with our previous work [10], we used the same but older perceptual model. More advanced ones can replace these ones to further the accuracy.

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

Table 1
Basic response events included in each assessment paradigm

Paradigm	Look	Point	Smile	Speak	Leave
RN	✓			✓	
SS	✓		✓		
IG	✓	✓			
RJA	✓				
IJA	✓	✓			
SA	✓				✓

3.2. Script Transcription Module

The behavioral logs integrate outputs from multiple models, producing heterogeneous data formats, such as body coordinates, eye vectors, facial expression types, gesture types, and speech content, etc.. To transform these behavioral logs into text that LLMs can understand, we developed a script transcription module to standardize and unify these diverse outputs into a coherent script.

3.2.1. *Response Parser*

Although the behavioral logs contain information about the actions of each individual in the video, they cannot be directly input into LLMs for analysis. There are two main reasons for this: 1. The perceptual model outputs either categories (e.g., facial expression categories) or values (e.g., gaze angles), and abstract inputs prevent LLMs from understanding the underlining behaviors. 2. Behavioral logs are recorded frame-by-frame, and lengthy inputs can lead to LLMs forgetting crucial information or struggling to detect ASD symptom representations.

To address these issues, we extract key response events to simplify the behavioral logs. In this paper, each video recording includes six paradigms, namely Response to Name (RN) [20], Social Smile (SS) [61], Indicating Gesture (IG) [104], Responding to Joint Attention (RJA) [78], Initiating Joint Attention (IJA) [51], and Separation Anxiety (SA) [10]. As shown in Table 1, We summarize the response events in the above paradigm that are most important for ASD observation. Generally, the observation of each paradigm is based on the following events. We define E as the paradigm's set of all doctor-patient interactions. E_1 represents the event in which the child looks at the target object, called Look at Object. E_2 represents the event in which the child points at the target object, called Point to Object. E_3 denotes the event where the child smiles, called Smile. E_4 denotes the event when the child or doctor speaks, called Speak. E_5 represents the event in which parents or doctors exit the testing studio, called Leave. The following formulation defines the response parsing process:

$$R = E([B]) \text{ w.r.t. } E = \left[E_i\right]_{i=1,2,\ldots 5} \tag{2}$$

where R records the timestamp of event E occurring in $[B]$.

Algorithm 1 Find Emotinonal Dynamic Points

Input: The Valence sequence $\left[a_1, a_2, \ldots, a_n\right]$ for children with ASD
output: The emotional dynamic points P_n , and its time interval $[t_{n\_start}, t_{n\_end}]$
1: Calculate the first-order derivatives of the valence sequence , and obtain the first-order derivative sequence $\left[d_1, d_2, \ldots, d_n\right]$.
2: **for** each $i \in [0, n]$ **do**
3: **if** $d_n > 0.2$ or $dn < 0.2$ **then**
4: Record the moment t_n of the P_n, and the interval of dynamic $[t_n - 0.5s, t_n + 0.5s]$.
5: **end if**
6: **end for**
7: **while** P_n and P_{n+1} have overlap **do**
8: Merged into one continuous emotional segment: $P_n = [t_{n\_start}, t_{n+1\_end}]$.
9: **end while**

3.2.2. *Response Textualization Module*

After capturing a response event, we describe its occurrence and behavior in the text format. The response textualization module can convert the predefined events in the paradigm video into textual descriptions. As shown in Fig. 4, this illustrates the textualization process for the Response to Name paradigm. In the paradigm, the child participant is first guided to play with toys on the desk. Once their attention is engaged, the assessor suddenly calls the child's name from behind. The child can exhibit one of three responses: 1) no response, 2) turning to face the caller, and 3) turning to face the caller and responding verbally. Based on instructions and responses, we describe the RN paradigm process using text. In addition, we found that adding background and gender descriptions before the instruction description can improve diagnostic accuracy.

3.2.3. *Emotion Textualization Module*

The DSM-V [67] emphasizes that "ASD children lack social or emotional reciprocity." This deficit manifests in their difficulty understanding the emotions of others, being unresponsive, or showing indifference to emotional signals. This could be visually reflected in facial expressions' valence or arousal values. However, the paradigm design in our dataset only considers category emotion classes (e.g., smile or neutral) in the Social Smile (SS) paradigm, as shown in Table 1. To enable LLMs to better understand the deficits in emotional reciprocity, we aim to reflect continuous emotional dynamics in the script. Therefore, we designed an emotion textualization module to capture the emotional dynamic points and transform video segments near these points into textual emotion descriptions.

The first-order derivative of emotion represents the rate of emotional dynamics and serves as a key feature for understanding emotional dynamics [75, 32, 103]. To capture emotional dynamic points, we approximate the degree of emotion variation using the first-order derivative of the

Script Description

System prompt

As an autism spectrum disorder diagnostic specialist, You know all the symptoms and criteria for judging Autism Spectrum Disorder(ASD). Now you will watch a video of three people interacting in a room. The descriptions of the videos are presented to you in temporal order, just as if you were watching them directly. You can disambiguate them in time. But you can't create a video plot out of thin air.

Script Content

01:03.625-01:04.625: The doctor called out the child's name.
01:04.500-01:05.875: The child turns alertly toward the doctor and look at the doctor normally. ……

Domain Prompt

I would like you to analyze the content of the video according to the following **[ASD symptoms].** The Preparatory phase is the pre-paradigm environment setting, and [ASD symptoms] in the preparation stage are not observed. You can only judge ASD when the most of following [ASD symptoms] occur simultaneously and repeatedly. Do not assume that the child focusing on the toy car suggests a highly restricted interest, and focus on the child's response after the doctor or parents give instructions.

[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 movements. 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.]

Format Prompt

In the end, Please judge whether the child is ASD or TD based on the content of the video, 1 means ASD, 0 means TD, and you can only answer 0 or 1 and please give the reason for the judgement.
1. Reasons for autism spectrum disorder judgment: [ ]
2. Autism spectrum disorder Judgment results: [ ]

Figure 5: The details of the script description. The blue part is human experience. The bolded portion is domain knowledge.

valence value. The detailed process of finding emotional dynamic points is described in Algorithm 1. Let $[a_1, a_2, \ldots, a_n]$ represent the valence sequence of facial expression, where a_n is the valence value of the nth frame. Similarly, let $[d_1, d_2, \ldots, d_n]$ denote the sequence of first-order derivatives of the valence sequence, where d_n represents the first-order derivative of the valence value at the n th frame. We define P as the emotional dynamic point and identify its location based on the emotional dynamic threshold α. Specifically, we identify a P_n when $d_n > 0.2$ or $d_n < -0.2$ and define the 1-second video segment before and after this point as a emotional dynamic segment. Notably, a paradigm video may contain multiple emotional dynamic segments. If these segments overlap, they are merged into a single continuous mood segment. Finally, we get the updated emotional dynamic points P and their time intervals $[t_{n\_start}, t_{n\_end}]$.

In transforming emotional fragments into emotional descriptions, it is challenging to manually summarize emotional dynamics into a limited set of discrete emotional events. In recent years, the general video understanding LLMs [72, 7, 76] has demonstrated a remarkable ability to generate video descriptions, enabling the transformation of video content into emotionally relevant textual descriptions. Given that our data includes both video and audio modalities, we employ the audio-visual LLM model (video-SALMONN [76]), to analyze visual frame sequences and audio events, with a primary focus on capturing emotionally relevant content. For segments, we generate textual descriptions of audio-visual content by prompting it with emotion-related queries. Figure 13 illustrates the emotion prompt and the emotion question. Fig 6 shows the complete process of the emotion textualization process. The valence line in the graph represents the emotional intensity of the child. This TD child responded by looking back at the doctor after hearing the command to call his name. The red dots on the reaction times indicate moments of large dynamics in valence. Clearly, the video shows the child's transition from a calm to a positive emotion.

3.2.4. Domain Prompt Module

Prompt engineering can enhance LLMs' understanding of script descriptions. Therefore, we design a domain prompt module for ASD detection, as detailed in Fig 5. The domain prompt module consists of four components: 1) The system prompt emphasizes the LLMs' identity to ensure they understand scripts in the temporal order. 2) The script content is generated by the response parser, the response textualization module and the emotion textualization. It records key events of the audio-visual behavior data in the text format as [timestamp, behavioral description]. 3) The domain prompt incorporates domain knowledge [41] and experience into the script descriptions. In this context, domain knowledge refers to the ASD diagnostic criteria, while experience represents the researcher's experiential knowledge in the clinical setup. 4) The format prompt constrains the output results.

3.3. Large Language Models

In this paper, we selected eight LLMs with strong performance and reputation: three closed-source models (GPT-4O [27], Claude 3.5-sonnet [2], Monnshot [50]) and five open-source models (LLAMA 3.1 [18], qwen2-72B-instruct [94], DeepSeek-R1-Distill-Qwen-32B [23], DeepSeek-R1-Distill-Llama-70B [23], DeepSeek-R1-671B [23]. We input script descriptions and user questions into these LLMs, and the assisted ASD detection results and interpretations were extracted from the answers. To ensure privacy, the script descriptions submitted to the closed-source model exclude any personal identifiers or physical attributes of the children. Verbal content is limited to broad descriptions of the physician's instructions and the child's corresponding responses, without disclosure of verbatim dialogue. Some representative script examples are presented here [2] .For few-shot learning, we only evaluate an open-source model deployed in our server because uploading data and label to a closed-source API may lead to data leakage. All LLMs are configured for fair experimentation with a *max_token* limit of 1000 and a *temperature* setting of 0.7.

3.4. The joint detection of multiple LLMs

To further improve detection accuracy and reduce the hallucination from individual LLM, we fuse the outputs of multiple LLMs. The most straightforward strategy is a

[2] https://github.com/lwx0724/Script-Centric-Behavior-Understanding

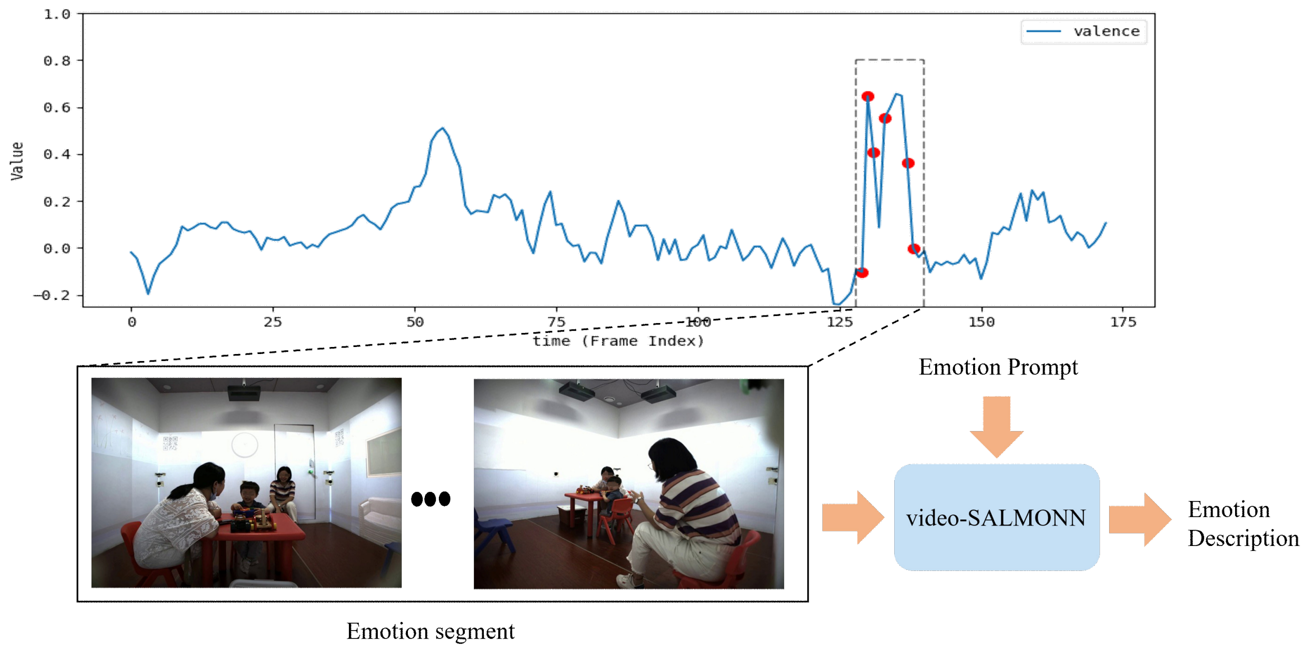


Figure 6: The emotion textualization process of the Response to Name paradigm. The blue line indicates the valence value. The red dotted line indicates the moment when the doctor calls child's name. The red dots indicate points of emotional dynamic, and the black dashed interval highlights the segment where the child responded.

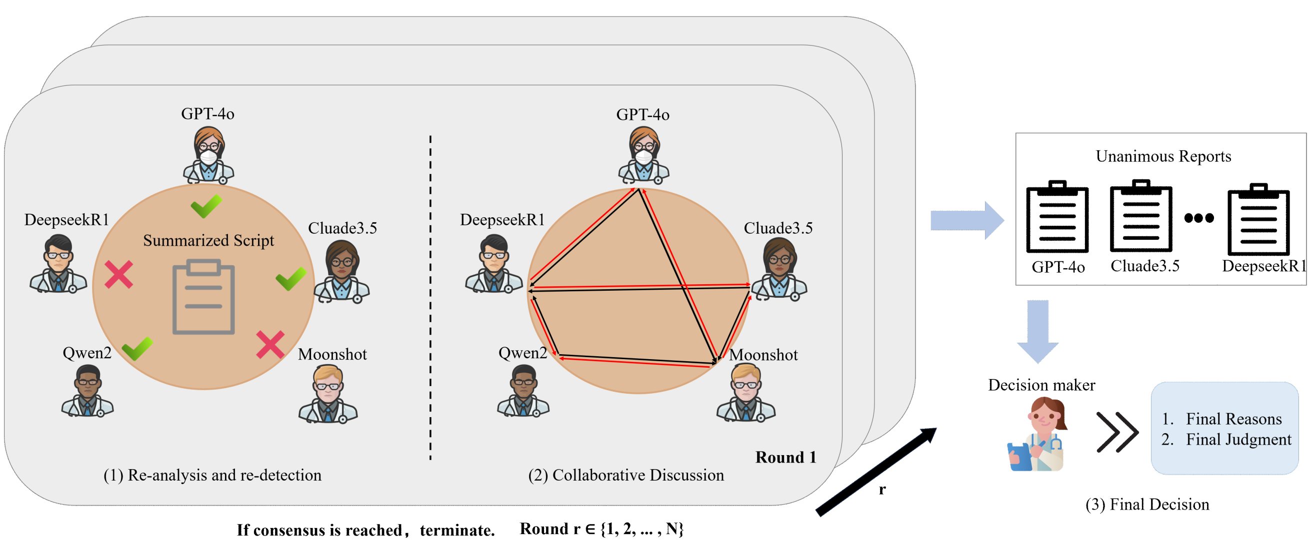
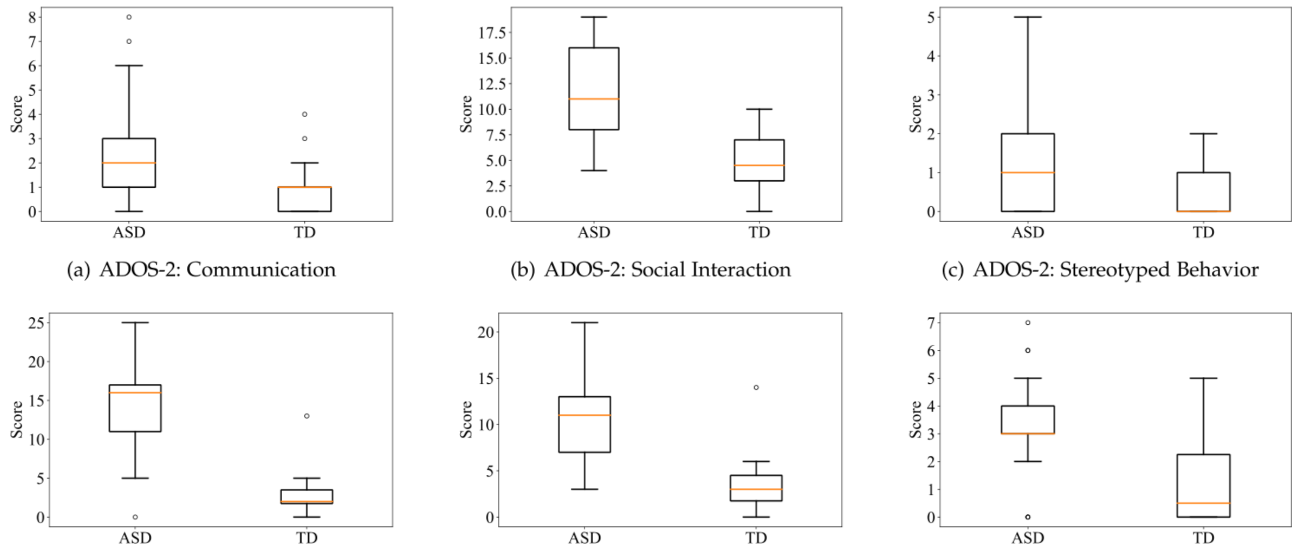


Figure 7: This is the framework for the joint detection of multiple LLMs. The process is divided into three stages: (1) Re-analysis and re-detection, (2) Collaborative Discussion, (3) Final Decision.

majority voting method, which we refer to as SCBU-Vote. We selecte five LLMs with high variability (GPT-4O [27], Claude 3.5-sonnet [2], Monnshot [50], qwen2-72B-instruct [94], DeepSeek-R1-671B [23]), and their predictions were aggregated through a voting mechanism to determine the final result. However, a clear limitation of this approach is that voting only applies to the final detection outcome, without providing a unified or interpretable rationale for the final decision.

Inspired by multi-agent medical diagnostics [81], we propose a framework for the joint detection of multiple LLMs called SCBU-Agents. The framework is divided into three stages, as shown in Fig 7: (1) Re-analysis and re-detection. Based on the behavioral scripts, the detection results of individual LLMs in the previous round, and the content of the discussion in the previous round, the behavioral scripts are reanalyzed, and the ASD detection results are re-evaluated. (2) Collaborative Discussion. Discuss the



(a) ADOS-2: Communication (b) ADOS-2: Social Interaction (c) ADOS-2: Stereotyped Behavior

(d) ADI-R: Reciprocal Social Interaction (e) ADI-R: Communication (f) ADI-R: Restricted Repetitive Behavior

Figure 8: Statistics of ASD/TD groups in the clinical database.The box stretches from the first quartile (Q1) to the third quartile (Q3) of each score distribution, with a colorful line representing the median. The whiskers extend the box to 1.5 times the interquartile range (IQR). Flier points denote scores beyond the end of the whiskers. (duplicated from Fig. 5 in [10])

behavioral scripts with other LLMs and try to convince other LLMs who have different opinions. If a consensus is reached or the maximum number of discussion rounds is exceeded, the process proceeds to Stage 3. Otherwise, it returns to Stage 1 for reanalysis. (3) Final Decision. We use GPT-4o to play the role of a decision maker and try to summarize the LLMs' detection results of last round and get the final results.

4. Experimental results

4.1. Dataset

We utilized the multimodal behavioral database in [10] to evaluate our method. This database comprises RGB, RGB-D and audio data recorded in a real clinical environment. Specifically, the dataset included eight distinct views of RGB HD video with a resolution of 4096 × 3000 and four different views of depth video with a resolution of 1280 × 720. These synchronized cameras ensured that the subject's behavior response were largely unobstructed.

The database included 95 participants, comprising 71 children diagnosed with ASD and 24 TD children. The TD group exhibited a more balanced distribution between males and females, whereas the ASD group had a significantly higher proportion of males compared to females. These characteristics align with the clinical distribution of the data , More details about this dataset can be found in [10]. For each assessment case, the physician will lead the child participant and parent through the six paradigms, which typically takes 20 to 30 minutes.

For diagnostic confirmation, all child participants underwent ADOS-2 [21] and ADI-R [44] assessments, followed by a comprehensive clinical evaluation conducted by multiple experienced clinicians, including at least one senior physician, to ensure diagnostic accuracy. Figure 5 presents the distribution of ADOS-2 and ADI-R scores across the different domains, illustrating the characteristic variations among participants from a clinical perspective.

All participants were Mandarin-speaking or Cantonese-speaking children from an urban region of China. Although specific socioeconomic indicators were not stored due to privacy regulations, the recruitment center predominantly serves middle-income families. Therefore, the dataset reflects a relatively homogeneous cultural and socioeconomic environment. The concise summary table of the dataset is shown in Table 2.

4.2. Baseline methods

We established comprehensive baseline comparisons, as shown in Table 3. Our experiments now encompass four categories of baselines. Our baseline experimental workflow and details are as follows.

(1) behavioral signal processing methods.

- Cheng et al. [10] extracts behavioral features from multiple signals, then detects ASD using machine learning algorithms.

(2) general video understanding methods

- video-salmonn [76] directly converts audio and video data into text, then uses an LLM to detect ASD.

(3) state-of-the-art general vision-language models.

- VideoCLIP [93] represents retrieval-style video-language alignment methods. We adopt VideoCLIP to extract a global video embedding for each clip and subsequently train a lightweight prediction head for ASD detection. Specifically, the video encoder is a 3D ConvNet (CNN architecture), producing features with a dimensionality of (1, 1024).

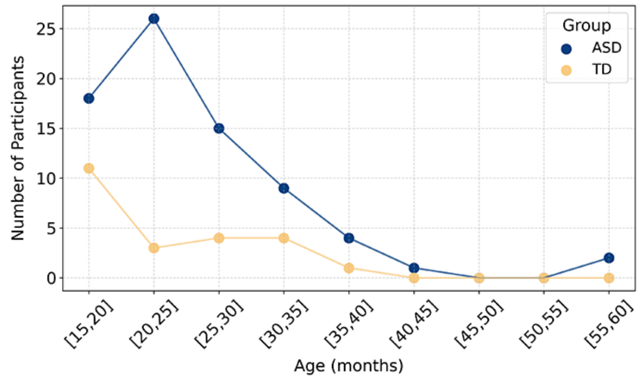
- OpenFlamingo-3B [6] represents Flamingo-style vision-language generative models. We input images and text (DSM-5 criteria) into OpenFlamingo-3B to extract multimodal features, then train a lightweight prediction head for ASD detection. Specifically, the image encoder is CLIP ViT-L/14 (Transformer architecture), yielding features of size (257, 1024), while the text encoder is MPT-1B, producing features of size (text-token-length, 2048). After cross-attention fusion, the resulting multimodal representation has a final dimensionality of (257, 1024).
- LLaVA-Med [34] represents a domain-adapted vision-language assistant trained with cross-modal instruction tuning. We uniformly sample video frames at 1 FPS and apply LLaVA-Med to perform image-level ASD detection. Specifically, we use the official checkpoint [3] and input each sampled frame together with a diagnostic query (e.g., "Please determine whether the child in the image has ASD and provide your reasoning.").

(4) deep pretraining model-based methods.

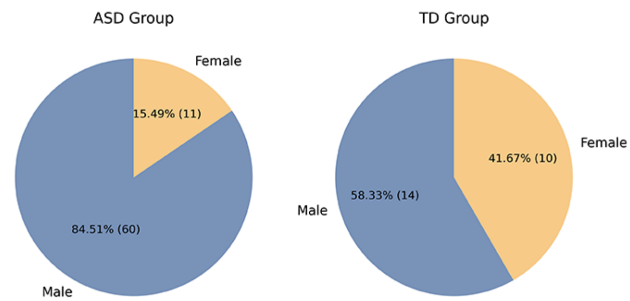
- Vision Transformer (VIT) [17] is used to extract visual representations, followed by training a prediction head for ASD detection. The extracted video features have a dimensionality of (1568, 1024) per second, and temporal aggregation is performed through average pooling along the time dimension.
- Whisper-Large-V3 [62] is used to extract speech/audio representations, after which a prediction head is trained for ASD detection. The extracted audio features have a dimensionality of (1500, 1280) per second, and temporal aggregation is conducted using average pooling along the time dimension.
- Multimodal Transformer (MulT) [83] is employed to perform temporal cross-modal fusion of video and audio features, followed by training a prediction head for ASD detection. To address the mismatch in feature dimensionality between modalities, we apply learnable linear projection layers before cross-modal fusion.

4.3. Overall Performance

We employ four widely used metrics to evaluate methods in ASD detection: Accuracy (ACC), F1-score (F1), Sensitivity (SN), and Specificity (SP). Accuracy represents the percentage of correct ASD predictions. The F1-score considers both recall and precision, reflecting the overall performance



(a) Age Distributions



(b) Gender Distributions

Figure 9: Statistics for ASD and TD groups in the clinical database. Scatterplot show age distribution of ASD and TD, pie chart show gender distribution of ASD and TD

Table 2
Summary of the clinical dataset used in this study.

Category	Details
Total participants	95 children
ASD / TD	74 ASD, 21 TD
Age range (months)	18–36 months
Mean age ± SD	24.87 ± 6.95
Clinical setting	Third Affiliated Hospital of SYSU
Interaction paradigms	RN, SS, IG, RJA, IJA, SA.
Recording duration	20-30 minutes
Modalities	RGB video, audio, scripts (textualized)
Video resolution	4096×3000 (8 views,8 FPS)
Audio sampling rate	16 kHz
Ethical approval	IRB [2018]02-196-01, [2020]02-118-03

of the methods. Sensitivity indicates the ability to identify ASD cases correctly, while specificity measures the ability to identify non-ASD cases correctly. Higher values for these metrics indicate better method performance. There are two validation scenarios: (1) Leave One Out Cross Validation (LOOCV). Each child participant was treated as one fold. In each iteration, video and audio from one child were held out for testing, and data from all remaining (N-1) children were used for training. (2) All test. Model parameters are

[3] https://github.com/microsoft/LLaVA-Med

not trained. Each child sample are inferred. Overall, the test set for both methods comprises all child samples.

All experiments were conducted on four NVIDIA RTX A6000 GPUs (48 GB) and a 36-core Intel Xeon CPU. For SCBU zero-shot LLM inference, the processing time was approximately 10 seconds per subject, depending on the number of script tokens. For SCBU-Agents, multi-round collaborative reasoning required 120-180 seconds per subject, depending on the number of discussion rounds and the number of participating models.

Table 3 compares the results of different LLMs and the impact of incorporating emotion descriptions. In addition, our proposed zero-shot method SCBU is also compared with multiple baseline methods as well. When using a single LLM without emotion descriptions, SCBU-Claude 3.5 (w/o emotion) achieves an F1-score of 91.55%, remarkably close to the supervised learning baseline (92.20%). Similarly, SCBU utilizing other LLMs also achieves comparable accuracy, demonstrating the feasibility of LLMs for ASD detection. When using a single LLM with emotional descriptions, SCBU-DeepSeekR1-4O (w/ emotion) achieves an F1-score of 94.04%, surpassing the supervised learning baseline. This result indicates that incorporating the emotional dynamic descriptions enhances the distinction between ASD and TD. Adding appropriate emotional descriptions in behavior scripts can further enhance LLMs' understanding of ASD.

Finally, we fused the detection results of multiple LLMs using two strategies: SCBU-Vote and SCBU-Agents. The results of these fusion methods are presented in Table 3. Both approaches outperform all single-LLM methods and significantly exceed the supervised baseline. Notably, SCBU-Agents (w/ emotion) achieved the highest F1-score of 95.24%. Further details of the SCBU-Agents' discussion process and detection rationale are available at here [2].

In addition to achieving high diagnostic accuracy, Fig. 10 presents two examples demonstrating the interpretability of LLMs in autism detection. As shown in Fig. 10 1), The LLMs' response include the reasons for the judgments and the judgment results. The reasons provided by the LLM for its judgments include competencies in social skills, responses to name-calling, smile interactions, stereotypical behaviors, and age factors. The bolded section on the left highlights the child's deficits in social interactions, which align with the DSM-V criteria for ASD diagnosis. As shown in Fig. 10 2), This explains the LLM's detection in TD children. The LLM also considers ASD characteristics such as social and communication skills, attention, eye contact, desire to share, and age. In summary, the LLM's interpretation of the results aligns with human judgment expectations and can serve as an alternative physician reference. The results show that our method can explain the causes and enhance the credibility of assisted detection.

4.4. The Clinical Reliability of LLM Explanations

To the best of our knowledge, no standardized quantitative metric currently exists for evaluating the clinical correctness of LLM explanations in ASD behavioral assessment, we adopted a human-expert-based evaluation protocol, which is widely used in explainable medical AI research [73, 79].

We conducted a structured expert-based evaluation, following common practices in explainable medical AI research. Two researcher (each with more than three years of ASD assessment experience) independently assessed the explanations produced by SCBU-Agents. Each explanation was evaluated across four clinically relevant dimensions using a 5-point Likert scale (1 = poor, 5 = excellent): (1) **Clinical correctness:** Consistency with DSM-5/ADOS diagnostic principles. (2) **Behavioral grounding:** Whether the reasoning is directly supported by observable behaviors in the script. (3) **Interpretability:** Logical coherence, clarity, and readability of the explanation. (4) **Diagnostic usefulness:** Whether the explanation provides actionable insight supporting ASD/TD judgment.

Specifically, the clinicians' ratings yielded the following average scores across the 95 participants in Table 5. To assess inter-rater reliability, we computed the quadratic weighted Cohen's kappa between the two annotators. The kappa values for all four evaluation dimensions exceeded 0.80, which indicates that the behavioral scores were relatively reliable and less influenced by subjective variability.

These results indicate that the explanations generated by SCBU-Agents are clinically reliable, well-grounded in observable behavior and useful for supporting human diagnosis.

4.5. Behavioral Distinction

To verify whether behavioral descriptions can effectively distinguish ASD, Table 4 demonstrates that all behaviors exhibited significant differences in response latency and response duration ($p < 0.05$). The mean delay in TD children looking at or pointing to the target object after the doctor's command was shorter than that in ASD children, while the mean duration was more prolonged. Furthermore, the mean latency in chasing ability was also shorter in TD children than in those with ASD. These findings align with ASD characteristics, including reduced response flexibility and diminished social attention.

4.6. Emotional Dynamic Distinction

Valence and arousal are the two core dimensions of emotion, forming the fundamental framework of emotional representation. Valence determines the overall direction of emotion (positive or negative), while arousal describes its intensity [33]. Tseng et al. [84] demonstrated that the range of emotional ratings in the ASD group was consistently limited. As shown in Fig. 11, both the valence and arousal ranges of the ASD group were significantly smaller than those of the TD group. Therefore, we utilized the differences in valence and arousal values between the ASD and TD groups to identify the emotional dynamic points. Furthermore, a t-test was conducted on our data to determine which measure showed greater statistical significance. Table 4 presents the significance analysis of valence and arousal. There was a less

Table 3
Comparison with Behavioral signal-processing methods, general video-understanding models, general vision-language models and deep pretraining model-based methods. V represents video modality, A represents audio modality, I represents image modality. T represents text modality, ST represents script text.

Model	Emotion	Modality	Verification	ACC(%)	F1(%)	SN(%)	SP(%)
Cheng et al. [10]	-	V+A	LOOCV	88.42	92.20	91.55	79.17
video-salmonn [76]	-	V+A+T	All test	74.74	85.54	100.00	0.00
VideoCLIP [93]	-	V	LOOCV	76.26	86.92	92.89	15.28
OpenFlamingo [6]	-	V+T	LOOCV	78.87	86.71	89.54,	48.93
LLaVA-Med [34]	-	I+T	All test	63.56	69.73	75.92	32.18
ViT [17]	-	V	LOOCV	72.22	83.67	95.19	4.29
Whisper-Large-v3 [62]	-	A	LOOCV	60.39	70.56	63.35	51.07
MulT [83]	-	V+A	LOOCV	75.75	81.58	71.89	87.13
SCBU-GPT4O	w/o	V+A+ST	All test	86.32	90.78	90.14	75.00
	w/	V+A+ST	All test	88.42	92.62	97.18	62.50
SCBU-Claude3.5	w/o	V+A+ST	All test	87.37	91.55	91.55	75.00
	w/	V+A+ST	All test	87.37	91.18	87.32	85.75
SCBU-Moonshot	w/o	V+A+ST	All test	81.05	87.67	90.14	54.17
	w/	V+A+ST	All test	84.21	88.55	81.69	91.67
SCBU-Llama3.1	w/o	V+A+ST	All test	82.11	88.89	95.77	41.67
	w/	V+A+ST	All test	84.21	90.45	**100.00**	37.50
SCBU-Qwen2	w/o	V+A+ST	All test	83.16	88.41	85.92	75.00
	w/	V+A+ST	All test	87.37	91.18	87.52	87.50
SCBU-DeepSeekR1-Distill-Qwen-32B	w/o	V+A+ST	All test	83.16	87.69	80.28	91.67
	w/	V+A+ST	All test	86.32	90.51	87.32	83.33
SCBU-DeepSeekR1-Distill-Llama-70B	w/o	V+A+ST	All test	85.26	90.67	98.77	54.17
	w/	V+A+ST	All test	87.37	92.21	100.00	50.00
SCBU-DeepSeekR1-671B	w/o	V+A+ST	All test	86.32	91.03	92.96	66.67
	w/	V+A+ST	All test	90.53	94.04	100.00	62.50
SCBU-OLMo2	w/o	V+A+ST	All test	83.16	89.74	98.59	37.50
	w/	V+A+ST	All test	84.21	90.00	88.73	70.83
SCBU-Vote	w/o	V+A+ST	All test	86.32	90.28	91.55	70.83
	w/	V+A+ST	All test	**92.63**	94.44	95.77	**88.33**
SCBU-Agents	w/o	V+A+ST	All test	87.37	91.78	94.37	66.67
	w/	V+A+ST	All test	**92.63**	**95.24**	98.59	75.00

significant difference in the Arousal minimum between the TD and ASD groups ($p = 0.1184$). Similarly, for the arousal maximum, the difference was also not high statistically significant ($p = 0.0583$). In contrast, the TD group had a wider range of valence than the ASD group, and both the valence maximum and minimum differed significantly ($p < 0.001$). Hence, this paper utilizes valence values to describe the emotional dynamic points in children with ASD.

Why do the emotional dynamic points contribute to distinguish TD from ASD? Differences in emotional dynamics are demonstrated in Table 4. Children with ASD exhibit a lower frequency of emotional dynamics compared to TD children, with a statistically significant difference ($p < 0.001$). Additionally, emotional dynamics in children with ASD occurred later than in TD children, with a significant difference in latency ($p < 0.05$).

4.7. Ablation Study

General video understanding model Since generic large video understanding models can generate video descriptions, could they also facilitate the generation of script descriptions? To investigate this, we design the following ablation experiments. First, we use the video-SALMONN [76] model to generate descriptions for each paradigm video. These descriptions were then chronologically integrated into script descriptions. Finally, we maintained the same domain prompts and LLMs when evaluating performance to ensure experimental fairness. As shown in Table 6, the scripts generated by the video understanding model showed a marked decline in performance across various tested LLMs. Notably, the specificity of all models approaches zero, indicating that they have lost the ability to classify TD cases accurately. As shown in Table 7, We conducted a

ASD

1. Reasons for autism spectrum disorder judgment: [The child displays several concerning behaviors consistent with ASD. **She shows limited social responsiveness**, failing to react to her name being called multiple times and not engaging in social smiling or reciprocal interactions. Her joint attention skills appear impaired, as she rarely looks at or points to objects when directed. The child demonstrates a fixated interest in the toy car, playing with it continuously throughout the session and showing resistance to shifting attention. **She exhibits minimal eye contact and doesn't initiate social interactions. The child's lack of response to separation from her parent is also atypical for her age.** While some behaviors are age-appropriate, the overall pattern suggests deficits in social communication and restricted, repetitive interests characteristic of ASD.]

2. Autism spectrum disorder Judgment results: [1]

1) Diagnostic Rationale for ASD

TD

1. Reasons for autism spectrum disorder judgment: [Based on the video content, the child demonstrates typical social and communication skills for his age. He responds promptly to his name, engages in joint attention, and shows appropriate social interaction. The child makes eye contact, shares findings with the doctor, and exhibits normal reactions to various stimuli. While there are instances where the child doesn't respond immediately, these are within normal variation for a 16-month-old. **The child shows no signs of repetitive behaviors, restricted interests, or sensory sensitivities**. His play with the toy car appears age-appropriate, **and he shows flexibility in shifting attention to new stimuli**. Overall, the child's behavior aligns more closely with typical development than with autism spectrum disorder.]

2. Autism spectrum disorder Judgment results: [0]

2) Diagnostic Rationale for TD

Figure 10: Examples of LLM detection for children with ASD and TD

Table 4
Comparison in valence, arousal, response Latency, response duration and dynamic frequency between the TD group and the ASD group

Dependent variable	TD	ASD	T Value	P Value
Response Latency				
look	4.9932	5.6257	-2.5202	0.0154
point	8.5822	10.2493	-3.0459	0.0042
chase	5.4036	10.0498	-2.1848	0.0321
Response Duration				
look	1.6503	1.3006	2.2138	0.0314
point	0.4355	0.2025	-2.1848	0.0506
chase	10.3385	16.2931	-5.6688	< 0.001
Valence Maximum	0.3294	0.2840	5.3101	< 0.001
Valence Minimum	-0.2905	-0.2540	-4.5054	< 0.001
Arousal Maximum	0.1977	0.1888	1.8850	0.0583
Arousal Minimum	-0.1715	-0.1648	-1.5625	0.1184
dynamic Frequency	9.1150	6.8513	3.5492	< 0.001
dynamic Latency	5.7320	6.7456	-2.1495	0.0319

1 The Response Latency, the Response Duaration and the dynamic Latency are measured in seconds.
2 The values of valence and arousal range from -1 to 1.
3 The dynamic Frequency is measured in times.

few-shot test using script descriptions generated by video-SALMONN, and the results similarly demonstrated the difficulty LLMs face in directly learning and analyzing from scripts generated by video understanding models. These experiments demonstrate the inability of LLMs to analyze scripts generated directly by video understanding models and underscores the importance of the transcription module we designed. A possible explanation for this is that behavior script descriptions capture critical responses of children with ASD during the paradigm process. In contrast, descriptions

Table 5
Clinical reliability scores of LLM-generated explanations (Likert 1–5). Mean and standard deviation of scores in the test dataset (N=95). The scoring rubric is listed in appendix C.

Metric	annotator 1	annotator 2	Kappa
Clinical correctness	4.32 ± 0.75	4.36 ± 0.62	0.86
Behavioral grounding	4.27 ± 0.87	4.36 ± 0.68	0.83
Interpretability	4.01 ± 0.86	4.11 ± 0.74	0.88
Diagnostic usefulness	4.06 ± 0.87	4.18 ± 0.71	0.84
clinical reliability (mean)	4.17	4.25	0.86

generated by generic video understanding models tend to be overly general, failing to capture the detailed discriminative behavior response.

Few-Shot Given the fairness and label leakage concerns associated with closed-source LLMs that upload data, we conducted few-shot experiments on our locally deployed open-source models Qwen2 and DeepseekR1. Qwen2 has a maximum input token limit of 128,000, which constrains the length of the input script. After testing, we set the maximum few-shot number of scripts without emotional descriptions to 20. Since adding emotional descriptions increases the length of a single script, the maximum few-shot number is set as 8. To ensure experimental fairness, we divided the 95 participants into two groups: A few-shot training set include 20 children, consisting of 10 ASD and 10 TD. The test set include 75 children, consisting of 61 ASD and 14 TD. As shown in Table 8, the evaluation metrics of SCBU-Qwen2 (w/o Emotion) continue to improve as the number of learning examples increases, ultimately reaching peak performance at a few-shot number of 20. It surpasses

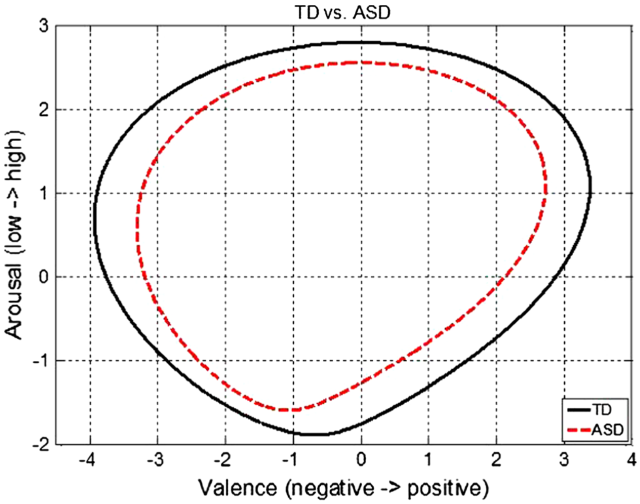


Figure 11: Comparison in the ranges of valence and arousal between the ASD group and the TD group (duplicated from Fig. 3 in [84])

the supervised learning baseline proposed by Cheng et al. [10] in all metrics. All experiments demonstrate that LLMs exhibit strong domain learning capabilities when provided with a few examples. Additionally, we conduct few-shot tests using DeepseekR1-Distill-Llama-70B (w/o Emotion) and DeepseekR1-Distill-Llama-70B (w Emotion). The results exhibit a trend consistent with Qwen2, with the best performance observed in the cases of DeepseekR1-Distill-Llama-70B(w/o Emotion, 20-shot) and DeepseekR1-Distill-Llama-70B (w Emotion, 8-shot). These results validate the effectiveness of few-shot learning and underscore our method's potential in scenarios with limited training data.

Emotion thresholds. The emotion threshold setting influences the number of emotional dynamic points. Specifically, a higher threshold results in fewer detected emotional dynamic points. If the number of emotional dynamics is too small, capturing variations in emotional valence becomes difficult. Conversely, an excessive number of dynamic points complicates the ability of LLMs to distinguish ASD from TD. Therefore, striking a balance between these factors requires careful selection of the emotion threshold. Fig 12 presents the evaluation metrics for different emotion thresholds.Thus, the emotion threshold for all experiments in this study is set to 0.175.

Emotion Description. Fig 13 provides an example of an emotion description. This description corresponds to the emotional dynamics in Fig 12. The emotion prompt makes LLM play the role of an ASD expert. The emotion question address several key aspects of ASD detection, such as eye contact, emotional expression, etc., ensuring a greater focus on emotionally relevant information related to ASD. The Description section in Fig 13 contains a textual description of the emotional dynamics observed in TD children. The bolded sections describe children's emotions, response sensitivity, and social initiative related to play, aligning with the actual events in the video. These suggest that the video-SALMONN model can describe emotion-related information.

Table 6
ASD detection results for zero-shot evaluations with script descriptions generated by video-SALMONN.

Method	ACC(%)	F1(%)	SN(%)	SP(%)
GPT4O	74.74	85.54	100.00	0.00
Claude3.5	75.79	85.89	98.59	8.33
Moonshot	74.74	85.54	100.00	0.00
Llama3.1	74.74	85.54	100.00	0.00
Qwen2	74.74	85.37	98.59	00.04
DeepSeekR1-32B	72.63	83.33	91.55	16.67
DeepSeekR1-70B	73.68	84.85	98.59	0.00
DeepSeekR1-671B	74.74	85.54	100.00	0.00

Table 7
ASD detection results for few-shot evaluations with script descriptions generated by video-SALMONN

Method	Emo	Few-Shot	ACC(%)	F1(%)	SN(%)	SP(%)
Qwen2	w/o	0-shot	72.00	83.72	98.18	0.00
		4-shot	73.33	84.62	100.00	0.00
		8-shot	76.00	85.94	100.00	10.00
		12-shot	74.67	85.27	100.00	5.00
		16-shot	76.00	85.94	100.00	10.00
		20-shot	77.33	86.61	100.00	15.00

Table 8
ASD detection results under the 20-training and 75-testing few-shot evaluation protocol. Emo donotes Emotion. FS donotes Few-Shot.

Method	Emo	FS	ACC	F1	SN	SP
cheng et al. [10]	-	20-shot	80.00	86.79	83.64	75.00
SCBU-Qwen2	w/o	0-shot	81.33	87.04	85.45	70.00
		4-shot	85.33	90.27	92.73	70.00
		8-shot	88.00	91.59	89.09	85.00
		12-shot	89.33	92.98	**96.36**	70.00
		16-shot	89.33	90.91	90.91	85.00
		20-shot	**92.00**	**94.34**	90.91	**95.00**
SCBU-Qwen2	w/	0-shot	85.33	87.27	87.27	80.00
		4-shot	85.67	89.72	87.27	80.00
		8-shot	**89.33**	**92.59**	**90.91**	**85.00**
SCBU-DeepSeekR1 -Distill-Llama-70B	w/o	0-shot	82.67	88.89	94.55	50.00
		4-shot	84.00	89.66	94.55	55.00
		8-shot	86.67	91.38	96.36	60.00
		12-shot	89.33	93.10	98.18	65.00
		16-shot	90.67	93.91	98.18	**70.00**
		20-shot	**90.67**	**94.02**	**100.00**	65.00
SCBU-DeepSeekR1 -Distill-Llama-70B	w/	0-shot	84.00	90.16	**100.00**	40.00
		4-shot	85.33	90.43	94.55	60.00
		8-shot	**89.33**	**92.98**	96.36	**70.00**

Table 9
ASD detection results under the leave-one-out cross validation protocol with SCBU-GPT4O. No denotes that no information has been added. DK denotes domain knowledge. HE denotes human experience. ED donotes emotion description.

No	DK	HE	ED	ACC(%)	F1(%)	SN(%)	SP(%)
✓				74.74	84.00	88.73	33.00
	✓			80.00	87.90	97.18	29.17
		✓		77.89	83.97	77.46	**79.17**
			✓	77.89	86.65	94.37	29.17
	✓	✓		84.21	89.36	88.73	70.83
	✓		✓	83.16	89.47	95.77	45.83
		✓	✓	85.26	90.00	88.73	75.00
	✓	✓	✓	**88.42**	**92.78**	**97.18**	62.50

Table 10
Comparison ASD detection results between our method and human expert assessments.

Method	ACC(%)	F1(%)	SN(%)	SP(%)
S_{anno1}	78.95	85.29	81.69	70.83
S_{anno2}	82.11	87.41	83.10	79.17
S_{vote}	83.16	88.41	85.92	75.00
SCBU-OLMo2	84.21	90.00	88.73	70.83
SCBU-Agents	**92.63**	**95.24**	**98.59**	**75.00**

Table 11
ASD detection results under the all test with SCBU-GPT4O. Point refers to gesture-related behaviors, Look refers to gaze-related behaviors, and Chase refers to running toward the parent.

point	look	chase	ACC(%)	F1(%)	SN(%)	SP(%)
✓			74.74	83.33	81.08	52.38
	✓		78.95	86.30	85.14	57.14
		✓	70.53	80.85	79.73	38.10
✓	✓		83.16	89.04	87.84	66.67
✓		✓	75.79	84.35	83.78	47.62
	✓	✓	80.00	87.02	86.49	57.14
✓	✓	✓	**88.42**	**92.78**	**97.18**	62.50

Domain Prompts Module. To demonstrate the validity, we examined the impact of different prior information by ablation experiments: domain knowledge from DSM-V, human experience from researchers and emotion description from emotional dynamic point. As shown in Table 9, the domain knowledge, the human experience, and emotion description can improve performance, respectively. Specifically, relying solely on domain knowledge can lead to high sensitivity but low specificity. However, incorporating human experience can effectively constrain the model, resulting in more balanced performance. Furthermore, adding emotional descriptions can comprehensively enhance the ASD detection capability of LLMs. These ablation experiments demonstrate the importance of designing tailored prompts for the ASD detection task. Our method achieved the best performance when all prior information are introduced simultaneously.

Behavioral module. To evaluate the contribution of each behavioral cue in the SCBU framework, we conduct an ablation study on three clinically meaningful behavior modules: Point (gesture-related behaviors), Look (gaze behaviors), and Chase (running toward the parent). For each module, we selectively enable or disable the corresponding behavior descriptions in the textual script and evaluate ASD detection performance using SCBU-GPT4O. The results are summarized in Table 11. When each module is used individually, performance drops substantially (70.53–78.95% ACC), indicating that no single behavioral cue is sufficient for reliable discrimination. Combining two modules leads to consistent improvements, with the Point and Look setting achieving the strongest two-module performance (83.16% ACC), reflecting the importance of integrating both gesture and gaze information for social-communication assessment. The best performance is obtained when all three behavior modules are included simultaneously, yielding an accuracy of 88.42%, an F1 score of 92.78%, and a sensitivity of 97.18%. These results demonstrate that each behavioral component contributes complementary diagnostic information and that comprehensive behavioral coverage significantly enhances ASD-TD classification within the SCBU framework.

4.8. comparion with human expert results

To compare the performance of our method against human clinicians using the same video data, two professionals with more than three years of clinical experience independently reviewed the recorded videos and assigned paradigm scores, denoted as $S_{anno1}$and S_{anno2}. Then, inconsistent labels will be retrospected to address the controversial annotations, resulting in the revised scoring results S_{vote}.

Based on physician experience, a total score34 was classified as ASD, while <34 indicated TD. Table 10 shows that the performance advantages of our method compared to human expert, validating its potential to support human diagnosis.

4.9. Fully-open large language model

Zero-shot in this work refers to a setting where LLMs perform diagnostic inference without any exposure to the target dataset or labeled examples during training or fine-tuning. The LLMs operated solely on textual behavioral scripts, without gradient-based updates or exposure to dataset-specific examples. This differs fundamentally from supervised methods such as Cheng et al [10] and end-to-end deep models [17, 62, 83], which trained classifiers on annotated ASD-TD pairs.

The domain-specific prompts provide clinical background knowledge (e.g., DSM-V criteria, behavioral definitions) but do not expose any data-dependent mappings between specific behaviors and outcomes in our dataset. Such prompting acts as knowledge conditioning rather than data-driven supervision, consistent with how zero-shot reasoning is defined in Med-PaLM [74] and DeepSeek-R1 [23]. Therefore, while our framework leverages structured prompts to guide reasoning, the detection process itself

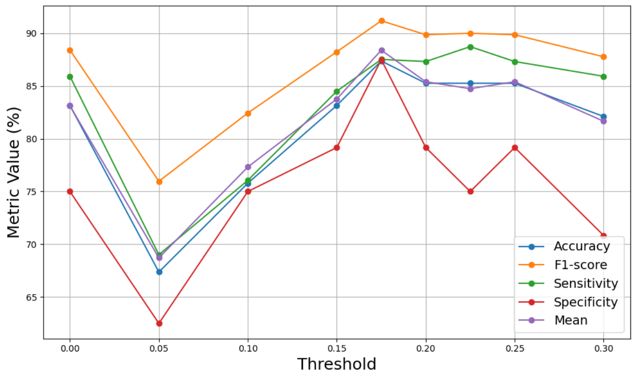


Figure 12: The impact of different emotional thresholds on different metric.

Emotion Description

Emotion Prompt

As an autism spectrum disorder diagnostic specialist, You know all the symptoms and criteria for judging Autism Spectrum Disorder(ASD).

Emotion Question

According to the child's reaction and emotion in the, sensitivitvideo, please describe this video from the child's eye contact, emotional expressiony to the reaction of others, and social initiative.

Description

From the perspective of the child in the video, it seems like a fun and engaging activity, sitting with adults and other children." The child's eye contact is focused on the board and the pieces being moved by the adult and other children. **The child's emotional expression was curiosity and interest as they watched the game and listened to the adults**. This child seems to be sensitive to the reactions of others as they look to adults and other children for clues as to what to do next. They also seem to enjoy the social aspects of the activity as they interact with adults and other children, saying "hello" and responding to adult cues. **Overall, the children seemed engaged, enjoyed the activity, were sensitive to the reactions of others, and showed social initiative by interacting with the adults and other children present.**

Figure 13: An example description of an emotional dynamic point.

remains zero-shot, meaning the models infer ASD likelihood based solely on their pre-trained reasoning capabilities.

To further verify this property, we additionally evaluated our framework using OLMo2 [55], a fully open LLM with publicly released corpora and parameters, confirming that no overlap exists with our ASD dataset. As shown in Table 10, both SCBU-OLMo2 (with emotion) and SCBU-OLMo2 (without emotion) surpass human experts' video evaluation in F1-score. These findings confirm that the proposed SCBU framework performs genuine zero-shot inference without relying on implicit supervision.

4.10. Limitations and Discussions

Although the sample size (N = 95) is relatively modest, it is consistent with or even larger than many prior studies in the autism research field [99, 42, 4]. Recruiting children with autism spectrum disorder requires strict diagnostic validation and ethical approval, making large-scale datasets challenging to obtain. This dataset employs diagnostic tools (ADOS [45], ADI-R [44]), IQ screening, and exclusion of major comorbidities to ensure data homogeneity. Furthermore, standardized data collection protocols guarantee signal reliability and minimize heterogeneity. Although Small sample sizes remain a persistent challenge in autism research, we will conduct long-term recruitment to expand our sample size, thereby further validating the generalizability of our methodology.

The cultural and socioeconomic homogeneity of the dataset may bias the behavioral patterns learned by the model. Social behaviors such as eye contact, gesture use, and emotional expressiveness vary across cultures, and differences in socioeconomic background may also influence child-adult interaction patterns. Thus, caution is required when generalizing the model to populations outside similar cultural or socioeconomic contexts. Future work will focus on multi-center, cross-cultural data collection to mitigate these biases.

Our method detects ASD by converting audio-visual behaviors into structured textual scripts. This conversion of clinical data into text helps prevent the leakage of patient privacy and facilitates data sharing among peers. Similarly, using text as a medium, our method adds descriptions of emotional dynamic points to the scripts, similar to adding valid handcrafted features to existing data. The textual format also enables clinicians' experiences to be translated into behavioral features, making it easier to model practical behavioral markers. Additionally, the LLMs analyze the patient's behavioral script, enabling the model to provide diagnostic explanations, which assist doctors in analyzing the causes of ASD in children.

Although our method performs well in ASD detection, it still has limitations in practical scenarios. First, the detection requires both audio and video to be recorded in a controlled environment, as it is essential to ensure that the quality of the data is sufficient for stable behavior analysis. Second, the diagnostic video content is limited to structured paradigms, and robust modules for behavior transcription are still not mature in unstructured settings. Finally, while emotional descriptions enhance the effectiveness of LLMs in ASD detection, they depend on the model's ability to understand audio-visual data. The current general video understanding models still have limitations in psychological analysis. In future work, we aim to enhance behavior perception and transcription in unstructured scenes or develop a generalized video understanding model focused on behavioral responses, enabling LLMs to better assess behavior in the ASD domain.

4.11. Ethics and Clinical Safety

The proposed system is intended solely as a clinician-assisted decision-support tool rather than an autonomous diagnostic instrument. Because false positives may introduce unnecessary anxiety or clinical risk, all outputs must be interpreted by licensed professionals. To protect privacy, all audiovisual data were collected under IRB approval, de-identified, and further transformed into anonymized behavioral scripts, avoiding the storage or processing of identifiable raw videos. The current framework has been evaluated only in controlled clinical settings and has not been validated for home-use, self-diagnosis, or deployment in uncontrolled environments. Any future clinical integration will require multi-site validation, regulatory review, and continuous clinician oversight to ensure safe and responsible use.

5. CONCLUSION

In this study, we propose a novel zero-shot and few-shot approach for detecting ASD using LLMs. We develop a script transcription Module to convert audio-visual content into text scripts. We design a domain prompts module to better leverage prior knowledge of ASD. Furthermore, we added an emotion textualization module to convert videos with intense emotional dynamics into textual descriptions, thereby enhancing behavioral representation. Extensive experimental results demonstrate the effectiveness of our method, showing strong zero-shot and few-shot capabilities. Moreover, LLMs explain the reasoning behind ASD detection, which helps physicians analyze the detection process more effectively. Future research will focus on developing either more straightforward and generalized behavioral perception models for video in unconstrained environments or general video understanding models in the ASD domain.

Algorithm 2 Script Transcription Module

Input: Behavior logs
output: Script Description

1: Response textualization module (section 3.2.2): Behavior logs are mapped to distinct responses, and the responses are transcribed into a behavioral description.
2: Segment the audio-video clips according to the timestamps of the emotional dynamic points P_n, and its time interval $[t_{n\_start}, t_{n\_end}]$.
3: Define a system prompt for the audio-video foundation model to specify the task and constrain the output format.
4: Formulate an emotion-description query that guides the model to generate affective interpretations of the segment.
5: Feed the segmented audio-video clip, system prompt, and emotion query into the multimodal model (Video-SALMONN) to obtain the corresponding emotional description.
6: Integrating behavioral description text and emotional description text to Script Description

A. The Algorithmic detail of the fusion strategy (SCBU-Agents)

The re-analysis and redetection stage recruits a group of ASD experts $AE = \{AE_1, AE_2, \ldots, AE_n\}$, where n represents the number of ASD experts. We assign different LLMs to distinct expert roles, each receiving domain prompts $Prompt_d$, behavioral scripts S, and the previous round's expert response RE_{r-1}, to generate the re-analyzed and re-detected report:

$$Repo = \mathrm{LLM}(S, RE_{r-1}, Prompt_d), \tag{3}$$

where r denotes the discussion round, and RE_0 indicates no discussion in the previous round.

The goal of the collaborative discussion phase is to explore behavioral scripts with other expert LLMs and to attempt to persuade other large language models with differing perspectives. Based on behavioral script S, its own diagnostic results and justification, and the diagnostic results and justification of expert AE_j, expert AE_i generates modification suggestions $Discussion_{ij}$ for AE_j:

$$Discussion_{ij} = AE_i(S, Repo_i, Repo_j). \tag{4}$$

The expert discussion result RE_r is represented as the set of all pairwise discussions:

$$RE_r = Discussion_{12} + Discussion_{13} + \ldots + Discussion_{(n-1)n}.$$

Table 12
Scoring criteria (1–5 scale) used by clinicians to assess the clinical reliability of LLM explanations. CC donotes Clinical correctness, BG donotes Behavioral grounding, IP donotes Interpretability, DU donotes Diagnostic usefulness.

Dimension	Score	Definition
CC	5	Fully consistent with DSM-5/ADOS-2 ASD characteristics; no clinical errors.
	4	Mostly consistent with diagnostic principles; only minor inaccuracies.
	3	Partially correct with noticeable omissions or inaccuracies.
	2	Mostly inconsistent with clinical knowledge.
	1	Completely incorrect or contradicts ASD/TD diagnostic criteria.
BG	5	All reasoning directly grounded in observable behaviors; no hallucinations.
	4	Mostly grounded with minimal speculative content.
	3	Mixed: some grounded reasoning, some speculative.
	2	Mostly speculative and not supported by observed behavior.
	1	Completely ungrounded or hallucinated explanations.
IP	5	Clear, coherent, and logically structured explanation.
	4	Generally understandable with minor issues.
	3	Understandable but contains redundancy or weak structure.
	2	Difficult to follow; unclear or poorly organized.
	1	Uninterpretable or lacks logical coherence.
DU	5	Provides strong, actionable cues directly supporting ASD/TD classification.
	4	Useful but not fully comprehensive.
	3	Limited diagnostic insight; partially helpful.
	2	Mostly irrelevant to diagnostic judgment.
	1	No diagnostic value or misleading.

Table 13
Instruction descriptions and response descriptions in the response textualization module

Paradigm	Instruction		Response	
	Session	description	name	description
RN	P1	The parent called out the child's name.	response 1	The child turns toward the doctor and look with saying hello.
	P2	The doctor called out the child's name.	response 2	The child turns toward the doctor and look .
			response 3	The child continued to play with the toy.
SS	P3	The doctor greets the child with a passional smile and say hello.	response 1	The child made eye contact with the doctor.
	P4	The doctor praises the child with a warm smile.	response 2	The child look at the doctor and smile .
	P5	The doctor plays a tickle game with smile. She slowly reaches out and gently touches the child.	response 3	The child smile but did not look at the doctor.
	P6	With a warm smile, parents entertain their children in whatever way they normally do in their daily lives.	response 4	The child bent his head and went on playing with the toy.
			response 5	The child made eye contact with the doctor without smile.
IG	P7	The doctor call the child's name and say "Look at that flower".	response 1	The child looked up in the direction of the picture.
	P8	The doctor call the child's name and say "Look at that tree".	response 2	The child keeps his head down and continues to play with his toy.
	P9	The doctor call the child's name and say "Look at that balloon"	response 3	The child precisely points out the location of the picture.
	p10	The doctor call the child's name and say "Look at that sofa"	response 4	The child roughly points out the location of the picture.
			response 5	The child turns around and makes eye contact with the doctor.
			response 6	The child keeps looking at the picture.
			response 7	Then the kid continue to play with his toy.
RJA	P11	The doctor raises his hand and points to the picture of a clock and says, "Look, there is a clock. what time it is.	response 1	The child turns his head backand then looks to the position of the clock.
			response 2	The child seek the clock while not finding the correct direction.
			response 3	The child looked up at the doctor's hand .
			response 4	The child keeps his head down and continues to play with his toy.
IJA	P12	The wall to the left of the child suddenly displays a yellow bird flapping its wings while a stereo plays the sound of bird.	response 1	The child is attracted to the animation playingand looks at the bird on the left wall.
	P13	The wall to the right of the child suddenly displays a moving riding car while the stereo plays the sound of the car moving.	response 2	The child turns around and makes eye contact with the doctor to share his findings.
	P14	A cow wiggling its ears is suddenly displayed on the wall behind the child's right side while the sound is played.	response 3	The child turns around and makes eye contact with the doctor to share his findings.
			response 4	The child keeps staring at the animation playing on the wall.
			response 5	The child raises his hand and points to the bird on the wall.
			response 6	The child lower his head again and continued to play with the toy.
SA	P15	The parent gets up from their seat, walks past the child, and finally leaves the room.	response 1	The child realizes that the parent has left and gets up and chases him toward the door.
	P16	The parent call the childś name outside the door and say, "Hi, mom is leaving. You have to play alone.	response 2	The child turns to the direction of the parent but remains seating at the table.
			response 3	The child keeps his head down and continues to play with his toy.
			response 4	The parents, the doctor and the child have left the room.
			response 5	The child lower his head again and continued to play with the toy.

(5)

When expert opinions fail to reach consensus or the maximum number of discussion rounds (set to 5 in this paper) has not been reached, the process returns to the first stage. Conversely, when collaborative discussions reach the maximum number of rounds or all experts reach consensus, the process enters the final decision stage.

At the final decision stage, we require the LLM to act as a medical decision-maker, deriving the final diagnosis based on behavioral scripts and expert reports:

$$FD = \mathrm{LLM}(S, Prompt_{dm}, Repo_1, \dots, Repo_n), \tag{6}$$

where FD denotes the final clinical diagnosis, $Prompt_{dm}$ represents the instructions for the decision-maker, and $Repo_n$ indicates the diagnostic report from the n-th expert.

B. Pseudocode for Script Transcription Module

Pseudocode for Script Transcription Module (script mapping + emotional textualization) such as algorithm 2.

The details of the Response textualization module, the system prompt, emotion-description query :

1. Response textualization module: Table 13 summarizes the textualized combinations generated based on the paradigm, the doctor's instructions, and the child's corresponding responses.

2. System prompt: As an autism spectrum disorder diagnostic specialist, you know all the symptoms and criteria for judging Autism Spectrum Disorder(ASD).

3. Emotion-description query: According to the child's reaction and emotion in the sensitivity video, please describe this video from the child's eye contact, emotional expression, to the reaction of others, and social initiative.

C. Explanation evaluation criteria

the evaluation criteria for each indicator are shown in Table 12.

References

[1] Aarthi, D., Kannimuthu, S., 2023. A comprehensive analysis of autism spectrum disorder using machine learning algorithms: Survey, in: International Conference on Power Engineering and Intelligent Systems (PEIS), Springer. pp. 241–253.

[2] Anthropic, 2024. Claude 3.5 sonnet. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet.

[3] Ardakani, H.A., Taghizadeh, M., Shayegh, F., 2022. Diagnosis of autism disorder based on deep network trained by augmented eeg signals. Int. J. Neural Syst. 32.

[4] Asmetha, J.R., Senthilkumar, R., 2025. Dsvtn-asd: Detection of stereotypical behaviors in individuals with autism spectrum disorder using a dual self-supervised video transformer network. Neurocomputing 624, 129397.

[5] Atyabi, A., et al., 2023. Stratification of children with autism spectrum disorder through fusion of temporal information in eye-gaze scan-paths. ACM Trans. Knowl. Discov. Data 17, 1–20.

[6] Awadalla, A., et al., 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 .

[7] Bai, S., et al., 2025. Qwen2.5-vl technical report. URL: https://arxiv.org/abs/2502.13923, doi:10.48550/ARXIV.2502.13923. *arXiv:2502.13923*.

[8] Begeer, S., Koot, H.M., Rieffe, C., Terwogt, M.M., Stegge, H., 2008. Emotional competence in children with autism: Diagnostic criteria and empirical evidence. Dev. Rev. 28, 342–369.

[9] Chen, S., Jiang, M., Zhao, Q., 2024. Deep learning to interpret autism spectrum disorder behind the camera. IEEE Transactions on Cognitive and Developmental Systems .

[10] Cheng, M., et al., 2023. Computer-aided autism spectrum disorder diagnosis with behavior signal processing. IEEE Trans. Affect. Comput. 14, 2982–3000.

[11] Chola Raja, K., Kannimuthu, S., 2023. Deep learning-based feature selection and prediction system for autism spectrum disorder using a hybrid meta-heuristics approach. Journal of Intelligent & Fuzzy Systems 45, 797–807.

[12] Colombo, M., Piccinini, G., 2023. The computational theory of mind. Cambridge University Press.

[13] Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S., 2020. Retinaface: Single-shot multi-level face localisation in the wild, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5203–5212.

[14] Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4690–4699.

[15] Deng, S., et al., 2024a. Hear me, see me, understand me: Audio-visual autism behavior recognition. IEEE Trans. Multimedia , 1–12Early access, doi: 10.1109/TMM.2024.3521838.

[16] Deng, S., et al., 2024b. Hear me, see me, understand me: Audio-visual autism behavior recognition. IEEE Trans. Multimedia , 1–12doi:10.1109/tmm.2024.3521838. early access, doi: 10.1109/TMM.2024.3521838.

[17] Dosovitskiy, A., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

[18] Dubey, A., et al., 2024. The llama 3 herd of models. doi:10.48550/arXiv.2407.21783. *arXiv:2407.21783*.

[19] Fan, R., Balaji Shankar, N., Alwan, A., 2024. Benchmarking children's asr with supervised and self-supervised speech foundation models, in: Proc. Annu. Conf. Int. Speech. Commun. Assoc. (INTERSPEECH), p. 5173–5177. doi:10.21437/Interspeech.2024-1353.

[20] Frohna, J.G., 2007. Failure to respond to name is indicator of possible autism spectrum disorder. J. Pediatrics 151, 327–328.

[21] Gotham, K., Risi, S., Pickles, A., Lord, C., 2007. The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. J. Autism Dev. Disord. 37, 613–627.

[22] Guevara, M., et al., 2024. Large language models to identify social determinants of health in electronic health records. NPJ digital medicine 7, 6.

[23] Guo, D., et al., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. doi:10.48550/ARXIV.2501.12948. *arXiv:2501.12948*.

[24] Hashemi, J., et al., 2021. Computer vision analysis for quantification of autism risk behaviors. IEEE Trans. Affect. Comput. 12, 215–226.

[25] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778.

[26] Huang, Z.A., Zhu, Z., Yau, C.H., Tan, K.C., 2021. Identifying autism spectrum disorder from resting-state fmri using deep belief network. IEEE Trans. Neural Netw. Learn. Syst. 32, 2847–2861.

[27] Hurst, A., et al., 2024. Gpt-4o system card. doi:10.48550/ARXIV.2410.21276. *arXiv:2410.21276*.

[28] Khan, K., Katarya, R., 2024. Aff-bpl: An adaptive feature fusion technique for the diagnosis of autism spectrum disorder using bat-pso-lstm based framework. Journal of Computational Science 83, 102447. doi:10.1016/j.jocs.2024.102447.

[29] Khan, K., Katarya, R., 2025a. Mcbert: A multi-modal framework for the diagnosis of autism spectrum disorder. Biological Psychology 194, 108976.

[30] Khan, K., Katarya, R., 2025b. Ws-bitm: Integrating white shark optimization with bi-lstm for enhanced autism spectrum disorder diagnosis. Journal of Neuroscience Methods 413, 110319. doi:10.1016/j.jneumeth.2024.110319.

[31] Kojovic, N., Natraj, S., Mohanty, S.P., Maillart, T., Schaer, M., 2021. Using 2d video-based pose estimation for automated prediction of autism spectrum disorders in young children. Sci Rep 11, 15069.

[32] Krone, T., Albers, C.J., Kuppens, P., Timmerman, M.E., 2018. A multivariate statistical model for emotion dynamics. Emotion 18, 739–754.

[33] Kuppens, P., Tuerlinckx, F., Russell, J.A., Barrett, L.F., 2013. The relation between valence and arousal in subjective experience. Psychol. Bull. 139, 917–940.

[34] Li, C., et al., 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, in: Adv. neural inf. proces. syst. (NeurIPS), pp. 28541–28564.

[35] Li, J., Chen, Z., Li, G., Ouyang, G., Li, X., 2022a. Automatic classification of asd children using appearance-based features from videos. Neurocomputing 470, 40–50.

[36] Li, J., Li, D., Savarese, S., Hoi, S., 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proc. Mach. Learn. Res. (ICML), pp. 19730–19742.

[37] Li, J., Zhong, Y., Han, J., Ouyang, G., Li, X., Liu, H., 2020. Classifying asd children with lstm based on raw videos. Neurocomputing 390, 226–238.

[38] Li, M., et al., 2019. An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. Comput. Speech Lang. 56, 80–94.

[39] Li, X., Zhang, D., Li, M., Lee, D.J., 2022b. Accurate head pose estimation using image rectification and a lightweight convolutional neural network. IEEE Trans. Multimedia 25, 2239–2251.

[40] Liu, R., Huang, Z.A., Hu, Y., Zhu, Z., Wong, K.C., Tan, K.C., 2024a. Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fmri data. IEEE Trans. Neural Netw. Learn. Syst. 35, 10591–10605.

[41] Liu, W., et al., 2023. Assessing the social skills of children with autism spectrum disorder via language-image pre-training models, in: Lect. Notes Comput. Sci. (PRCV), pp. 260–271.

[42] Liu, W., et al., 2024b. Self-guided dmt: exploring a novel paradigm of dance movement therapy in mixed reality for children with asd. IEEE Trans. Vis. Comput. Graph. 30, 2119–2128.

[43] Lord, C., Elsabbagh, M., Baird, G., Veenstra-Vanderweele, J., 2018. Autism spectrum disorder. The lancet 392, 508–520.

[44] Lord, C., Rutter, M., Le Couteur, A., 1994. Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J. Autism Dev. Disord. 24, 659–685.

[45] Lord, C., et al., 1989. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. J. Autism Dev. Disord. 19, 185–212.

[46] Lord, C., et al., 2022. The lancet commission on the future of care and clinical research in autism. The Lancet 399, 271–334.

[47] Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification, in: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops (CVPRW).

[48] Mazefsky, C.A., et al., 2013. The role of emotion regulation in autism spectrum disorder. J. Am. Acad. Child Adolesc. Psychiatr. 52, 679–688.

[49] McAdams, D.P., 2006. The role of narrative in personality psychology today. Narrative inquiry 16, 11–18.

[50] Moonshot, 2024. kimi. [Online]. Available: https://kimi.moonshot.cn/.

[51] Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A., Seibert, J., 2003. Early social communication scales (ESCS). [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/103595277/escs_manual_2003_2013-libre.pdf.

[52] Negin, F., Ozyer, B., Agahian, S., Kacdioglu, S., Ozyer, G.T., 2021. Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. Neurocomputing 446, 145–155.

[53] Neo, W.S., Foti, D., Keehn, B., Kelleher, B., 2023. Resting-state eeg power differences in autism spectrum disorder: a systematic review and meta-analysis. Transl. Psychiatr. 13, 389.

[54] Nori, H., et al., 2022. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. doi:10.48550/ARXIV.2311.16452. *arXiv:2311.16452*.

[55] OLMo, T., , et al., 2024. 2 olmo 2 furious. arXiv preprint arXiv:2501.00656 .

[56] Omiye, J.A., Gui, H., Rezaei, S.J., Zou, J., Daneshjou, R., 2024. Large language models in medicine: The potentials and pitfalls: A narrative review. Ann. Intern. Med. 177, 210–220.

[57] Park, K.W., Cho, S.B., 2023. A residual graph convolutional network with spatio-temporal features for autism classification from fmri brain images. Appl. Soft. Comput. 142, 110363.

[58] Piana, S., Malagoli, C., Usai, M.C., Camurri, A., 2021. Effects of computerized emotional training on children with high functioning autism. IEEE Trans. Affect. Comput. 12, 1045–1054.

[59] Povey, D., et al., 2011. The Kaldi speech recognition toolkit, in: IEEE Workshop Autom. Speech Recogn. Underst. (ASRU).

[60] Prakash, V.G., et al., 2023. Computer vision-based assessment of autistic children: Analyzing interactions, emotions, human pose, and life skills. IEEE Access 11, 47907–47929.

[61] Qiu, N., et al., 2020. Application of the still-face paradigm in early screening for high-risk autism spectrum disorder in infants and toddlers. Front. Pediatr. 8, 290.

[62] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision, in: Proc. Mach. Learn. Res. (ICML), pp. 28492–28518.

[63] Raja, K.C., Kannimuthu, S., 2023. Conditional generative adversarial network approach for autism prediction. Computer Systems Science & Engineering 44.

[64] Rani, A., Verma, Y., 2024. Activity-based early autism diagnosis using a multi-dataset supervised contrastive learning approach, in: Proc.- IEEE Winter Conf. Appl. Comput. Vis. (WACV), pp. 7788–7797.

[65] Rashidan, M.A., et al., 2021. Technology-assisted emotion recognition for autism spectrum disorder (ASD) children: a systematic literature review. IEEE Access 9, 33638–33653.

[66] Rashidan, M.A., et al., 2023. Stimuli video quantification based on valence-arousal elicitation in children with autism spectrum disorder (asd). IEEE Access 11, 111142–111153.

[67] Regier, D.A., Kuhl, E.A., Kupfer, D.J., 2013. The dsm-5: Classification and criteria changes. World psychiatry 12, 92–98.

[68] Reichow, B., 2012. Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders. J. Autism Dev. Disord. 42, 512–520.

[69] S, J.B., Pandian, D., Rajagopalan, S.S., Jayagopi, D., 2022. Detecting a child's stimming behaviours for autism spectrum disorder diagnosis using rgbpose-slowfast network, in: Proc. Int. Conf. Image Process. (ICIP), pp. 3356–3360.

[70] Sarabadani, S., Schudlo, L.C., Samadani, A.A., Kushski, A., 2020. Physiological detection of affective states in children with autism spectrum disorder. IEEE Trans. Affect. Comput. 11, 588–600.

[71] Schank, R.C., Abelson, R.P., 2013. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Psychology press.

[72] Shu, Y., et al., 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. doi:10.48550/ARXIV.2409.14485. *arXiv:2409.14485*.

[73] Singhal, K., et al., 2023a. Large language models encode clinical knowledge. Nature 620, 172–180.

[74] Singhal, K., et al., 2023b. Large language models encode clinical knowledge. Nature 620, 172–180.

[75] Steephen, J.E., 2013. Hed: A computational model of affective adaptation and emotion dynamics. IEEE Trans. Affect. Comput. 4, 197–210.

[76] Sun, G., et al., 2024. video-salmonn: Speech-enhanced audio-visual large language models. arXiv preprint arXiv:2406.15704 .

[77] Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5693–5703.

[78] Talbott, M.R., et al., 2020. Brief report: Preliminary feasibility of the tedi: A novel parent-administered telehealth assessment for autism spectrum disorder symptoms in the first year of life. J. Autism Dev. Disord. 50, 3432–3439.

[79] Tam, T.Y.C., et al., 2024. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ digital medicine 7, 258.

[80] Tang, C., Zheng, W., Zong, Y., Qiu, N., Lu, C., Zhang, X., Ke, X., Guan, C., 2020. Automatic identification of high-risk autism spectrum disorder: A feasibility study using video and audio data under the still-face paradigm. ieee transactions on neural systems and rehabilitation engineering 28, 2401–2410.

[81] Tang, X., et al., 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning, in: Find. Assoc. Comput Linguist. (ACL), p. 599–621.

[82] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. Nat. Med. 29, 1930–1940.

[83] Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences, in: Proc. Annu. Meet. Assoc. Comput Linguist. (ACL), p. 6558.

[84] Tseng, A., et al., 2013. Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. J. Autism Dev. Disord. 44, 1332–1346.

[85] Uddin, M.Z., Shahriar, M.A., Mahamood, M.N., Alnajjar, F., Pramanik, M.I., Ahad, M.A.R., 2024. Deep learning with image-based autism spectrum disorder analysis: A systematic review. Eng. Appl. Artif. Intell. 127, 107185.

[86] Wang, H., et al., 2023. Huatuo: Tuning llama model with chinese medical knowledge. doi:10.48550/ARXIV.2304.06975. *arXiv:2304.06975*.

[87] Wang, X., Zhang, R., Kong, T., Li, L., Shen, C., 2020. Solov2: Dynamic and fast instance segmentation, in: Adv. neural inf. proces. syst. (NeurIPS), pp. 17721–17732.

[88] Wei, P., Ahmedt-Aristizabal, D., Gammulle, H., Denman, S., Armin, M.A., 2023. Vision-based activity recognition in children with autism-related behaviors. Heliyon 9, 16763.

[89] White, J., et al., 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. doi:10.48550/ARXIV.2302.11382. *arXiv:2302.11382*.

[90] Wilkinson, C.L., et al., 2024. Developmental trajectories of eeg aperiodic and periodic components in children 2–44 months of age. Nat. Commun. 15, 5788.

[91] Wu, C., et al., 2023. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. arXiv

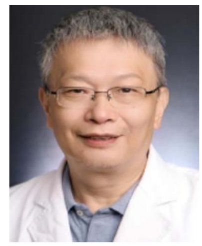
preprint arXiv:2310.09909 .
[92] Xia, C., Chen, H., Han, J., Zhang, D., Li, K., 2025. Identifying children with autism spectrum disorder via transformer-based representation learning from dynamic facial cues. IEEE Trans. Affect. Comput. 16, 83–97.
[93] Xu, H., et al., 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 .
[94] Yang, A., et al., 2024. Qwen2 technical report. doi:10.48550/ARXIV.2407.10671. *arXiv:2407.10671*.
[95] Yu, C., Zhang, D., Zou, W., Li, M., 2024. Joint training on multiple datasets with inconsistent labeling criteria for facial expression recognition. IEEE Trans. Affect. Comput. 15, 1812–1825.
[96] Zeng, A., et al., 2022. Glm-130b: An open bilingual pre-trained model. doi:10.48550/ARXIV.2210.02414. *arXiv:2210.02414*.
[97] Zhang, H., Li, X., Bing, L., 2023a. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. doi:10.48550/ARXIV.2306.02858. *arXiv:2306.02858*.
[98] Zhang, K., et al., 2024. A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine , 1–13.
[99] Zhang, N., Ruan, M., Wang, S., Paul, L., Li, X., 2023b. Discriminative few shot learning of facial dynamics in interview videos for autism trait classification. IEEE Trans. Affect. Comput. 14, 1110–1124.
[100] Zhang, N., Ruan, M., Wang, S., Paul, L., Li, X., 2023c. Discriminative few shot learning of facial dynamics in interview videos for autism trait classification. IEEE Trans. Affect. Comput. 14, 1110–1124.
[101] Zhou, H., et al., 2023. A survey of large language models in medicine: Progress, application, and challenge. doi:10.48550/ARXIV.2311.05112. *arXiv:2311.05112*.
[102] Zhou, W., Yang, M., Tang, J., Wang, J., Hu, B., 2024. Gaze patterns in children with autism spectrum disorder to emotional faces: Scanpath and similarity. IEEE Trans. Neural Syst. Rehabil. Eng. 32, 865–874. doi:10.1109/tnsre.2024.3361935.
[103] Zhu, X., et al., 2024. Daily life affective dynamics as transdiagnostic predictors of mental health symptoms: An ecological momentary assessment study. J. Affect. Disord. 351, 808–817.
[104] Zwaigenbaum, L., Bryson, S., Garon, N., 2013. Early identification of autism spectrum disorders. Behav. Brain Res. 251, 133–146.

Wenxing Liu received the master's degree in computer science from the Chongqing University of Technology. He is currently working toward the PhD degree in computer science with Wuhan University. His research interests include ASD diagnosis, ASD assessment, and multimodal large language model.

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

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

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



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

Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in 2013. He is currently a Professor of Electronical and Computer Engineering at Duke Kunshan University. He is also an Adjunct Professor at School of Computer Science in Wuhan University. His research interests are in the areas of audio, speech and language processing as well as multimodal behavior signal processing. He has published more than 200 papers and served as the member of IEEE speech and language technical committee, APSIPA speech and language processing technical committee, the editorial board member of the IEEE/ACM Transactions on Audio, Speech, and Language Processing and Computer Speech & Language. He is an area chair at Interspeech 2016, 2018, 2020 and 2024, 2025 as well as the technical program co-chair of Odyssey 2022 and ASRU 2023. Works co-authored with his colleagues have won first prize awards at Interspeech Computational Paralinguistic Challenges 2011, 2012 and 2019, ASRU 2019 MGB-5 ADI Challenge, Interspeech 2020 and 2021 Fearless Steps Challenges, VoxSRC 2021, 2022 and 2023 Challenges, ICASSP 2022 M2MeT Challenge, IJCAI 2023 ADD challenge, ICME 2024 ChatCLR challenge and Interspeech 2024 AVSE challenge. He received the IBM faculty award in 2016, the ISCA Computer Speech and Language 5-years best journal paper award in 2018 and the youth achievement award of outstanding scientific research achievements of Chinese higher education in 2020.