Neurocomputing 706 (2026) 135024

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Multimodal large language models for ADOS-M1 behavioral assessment☆,☆☆

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

[a] School of Computer Science, Wuhan University, Wuhan, 430072, Hubei, China
[b] Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, 510630, Guangdong, China
[c] Digital Innovation Research Center, Duke Kunshan University, 518172, Kunshan, Jiangsu, China
[d] School of Artificial Intelligence, The Chinese University of Hong Kong, 215316, Shenzhen, Guangdong, China
[e] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510006, Guangdong, China

## HIGHLIGHTS

- A large-scale ADOS-M1 dataset with refined paradigm annotations and item-level clinical scores.
- A multimodal framework integrates visual, audio, and dialogue cues through instruction-tuned LLMs.
- ADOS paradigm recognition and item-level behavior assessment are reformulated as structured text generation tasks.
- Accurate item-level prediction enables reliable subject-level ASD detection.
- Robust item-level behavior prediction from raw clinical recordings.

## ARTICLE INFO

Communicated by D. Liu

*Keywords:*
Autism spectrum disorder
ADOS behavioral assessment
Multimodal large language model

## ABSTRACT

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by deficits in social communication and interaction. The Autism Diagnostic Observation Schedule (ADOS) is widely regarded as the gold-standard clinical tool for ASD assessment; however, the need for extensive expert observation limits its scalability for large-scale screening. Recent studies in artificial intelligence have attempted to detect ASD from multimodal behavioral data, yet most approaches directly predict diagnostic labels and fail to model the structured scoring protocol used in ADOS evaluations. In this study, we propose a multimodal large language model framework for ADOS-M1 behavioral assessment from audio–visual data. The proposed method integrates visual, audio, and dialogue information through three modality branches. Lightweight adapters and Low-Rank Adaptation (LoRA) are used to align multimodal representations with the embedding space of instruction-tuned large language models (LLMs). Instead of predicting diagnostic labels directly, we reformulate ADOS paradigm recognition and item-level behavioral assessment as structured text generation tasks, enabling LLMs to reason over multimodal behavioral cues in accordance with clinical assessment rules. Experiments on a clinical ADOS-M1 dataset containing 447 participants demonstrate that the proposed framework significantly outperforms unimodal and multimodal baselines. The best model achieves 73.77% accuracy in paradigm classification and 72.61% mean accuracy averaged across all ADOS items at clip level, and enables subject-level ASD detection with up to 86.67% accuracy. These results demonstrate the effectiveness of multimodal LLMs for ASD behavioral assessment.

☆ This research is funded in part by the Guangdong Science and Technology Plan (2023A1111120012), and National Natural Science Foundation of China (62571223, 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).
* Corresponding author at: School of Artificial Intelligence, The Chinese University of Hong Kong, 215316, Shenzhen, Guangdong, China.
*Email address:* ming.li369@cuhk.edu.cn (M. Li).

https://doi.org/10.1016/j.neucom.2026.135024
Received 3 June 2026; Received in revised form 30 July 2026; Accepted 30 August 2026
Available online 3 September 2026
0925-2312/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

## 1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by persistent deficits in social communication and interaction skills, along with restricted and repetitive patterns of behavior [23]. There is broad consensus among experts that early and accurate diagnosis is essential for timely intervention and improved developmental outcomes [16,42,49]. Data reported by the U.S. Centers for Disease Control and Prevention indicate a significant and sustained increase in the prevalence of ASD in recent years [45]. Due to pronounced differences in social communication and interaction, autistic children often require long-term support. These neurodevelopmental traits can place substantial psychological and financial demands on families and present broader societal challenges [44].

Early and accurate diagnosis is essential for enabling timely intervention and improving developmental outcomes. The ADOS [43] is widely recognized as the gold standard for ASD diagnosis. This assessment evaluates multidimensional competencies such as language, social interaction, play, and restricted or repetitive behaviors through structured and semi-structured interaction paradigms. However, the ADOS assessment requires extensive expert observation and relies heavily on subjective behavioral interpretation, which limits its scalability and hinders widespread adoption, particularly in regions with limited healthcare infrastructure.

Therefore, there is an urgent need for objective and efficient screening tools that can be easily implemented on a large scale. To this end, researchers have increasingly investigated the application of artificial intelligence techniques for the automated analysis of autism-related behaviors using audio-visual data. Previous studies have investigated feature-based deep learning methods that train classifiers on visual [30,36,38,54,62,63], audio [5,28,34,46,71], text [3,11,41,58] or multimodal fusion [48,57,60] representations for autism detection and behavioral assessment. Despite their promising performance, these approaches exhibit three major limitations. First, unimodal representations are insufficient to capture the complex interactions among visual behaviors, speech prosody, and conversational semantics that are critical for ADOS behavioral assessment [19,47,48]. Second, conventional multimodal fusion frameworks typically follow a feature-level classification pipeline, lacking the capacity to reason over clinical scoring rules. Third, most existing models directly predict diagnostic labels without explicitly modeling fine-grained ADOS behavioral items, thereby limiting interpretability and clinical compatibility. Script-centric behavior understanding methods [37] attempt to address these issues by transforming audio–visual data into structured behavioral scripts and leveraging large language models for downstream reasoning. However, the effectiveness of behavioral script generation depends on standardized experimental settings and high-precision perception models. When generic video-language models are employed, substantial performance degradation is often observed [37].

Large language models (LLMs) [15,21,69] have recently demonstrated remarkable reasoning capabilities and strong generalization performance in instruction-following tasks, making them a promising direction for medical diagnosis [18,50,59]. However, substantial distributional discrepancies between perception models and language models make the direct application of LLMs to raw audio-visual signals challenging. Moreover, existing video-language models [56,65] remain limited in capturing the fine-grained behavioral patterns associated with ASD.

To address these challenges, we propose a multimodal framework that enables instruction-tuned LLMs to perform ADOS paradigm classification and item-level behavioral assessment. Specifically, we design three modal branches (visual, audio, and dialogue) to transform raw multimodal signals into token representations aligned with the LLM's embedding space. A lightweight adapter mechanism is introduced to bridge cross-modal discrepancies, incorporating LoRA [24] for parameter-efficient multimodal instruction tuning. Unlike conventional end-to-end classifiers, our framework reformulates both paradigm classification and item-level behavior assessment as structured text generation tasks, thereby allowing LLMs to reason in accordance with clinical rules. The methodological contribution of this work lies in formulating the structured ADOS-M1 protocol as an item-oriented multimodal generation task for clinically interpretable behavioral assessment.

We collected a large-scale clinical ADOS-M1 dataset comprising 447 participants and conducted extensive experiments to evaluate the effectiveness of the proposed framework. Compared with unimodal baselines, conventional multimodal fusion methods, and script-centric behavioral understanding methods, our approach achieves substantial improvements in both paradigm classification and item-level assessment tasks. Moreover, the predicted ADOS item scores effectively support subject-level ASD detection under both ADOS rule-based and machine learning backend based diagnostic protocols, highlighting the clinical relevance of fine-grained behavioral assessment.

The main contributions of this study are summarized as follows:

- We collect a large-scale ADOS-M1 clinical dataset with refined paradigm annotations, standardized item-level behavioral scores and expert diagnostic outcomes.
- We propose a multimodal large language model framework that integrates visual, audio, and dialogue information for automated ADOS paradigm recognition and behavioral assessment.
- We reformulate ADOS paradigm recognition and item-level behavioral assessment as structured text generation tasks and introduce multimodal instruction tuning with LoRA to adapt LLMs for clinical behavioral reasoning.
- Extensive experiments demonstrate that the proposed framework significantly improves paradigm classification and item-level behavioral assessment, enabling reliable subject-level ASD detection.

## 2. Related work

### 2.1. Feature-based methods for automated autism detection

Recent advances in artificial intelligence have promoted automated ASD detection from multimodal behavioral data. Early studies primarily focus on unimodal representations. Vision-based approaches model gaze patterns, body movement, or facial expressions using convolutional or transformer-based architectures. For example, Liu et al. [36] develop a machine learning framework based on eye-tracking features of face processing abnormalities to automatically identify ASD in children. Liu et al. [38] also introduce a multimodal dataset centered on response-to-name behavior and propose a machine learning framework that integrates audio and visual cues for automatic recognition of autism-related behaviors. Li et al. [30] propose an appearance-based gaze estimation method that extracts the 3D gaze direction from raw video to characterize gaze behavior and achieve high accuracy in detecting ASD. Wei et al. [62] propose a three-dimensional convolutional neural network that leverages spatiotemporal action features for the automatic detection of stereotyped movements in ASD. Chen et al. [63] 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. Audio-based methods enable early screening by analyzing vocalization modes, acoustic characteristics and intonation patterns. For example, Li et al. [34] present an automated assessment framework that integrates prosodic and textual features to identify atypical language behaviors in ASD. Mohanta and Mittal [46] investigate a feature-based pipeline that analyzes low-level acoustic speech descriptors and trains machine learning classifiers for ASD detection. Briend et al. [5] show that traditional acoustic features extracted from speech recordings can achieve high accuracy in ASD classification. Yin et al. [71] develop a machine learning-based early prediction framework that leverages acoustic features extracted from infants' vocalizations to identify ASD risk. In contrast, text-based methods utilize transcribed dialogues to capture linguistic and pragmatic anomalies. For example, Cho et al. [11] employ both textual and acoustic features

extracted from conversational data and demonstrate the effectiveness of machine learning models for ASD detection. Londhe et al. [41] combine NLP feature extraction methods such as Bag-of-Words, TF-IDF, and word embeddings with multiple machine learning approaches to validate the effectiveness of text-based ASD detection. Themistocleous et al. [58] integrate BERT [13] embeddings with machine learning models to classify children with ASD from typically developing (TD) based on narrative production and expressive vocabulary. Assaf et al. [3] develop a transcript-based ASD screening framework that extracts quantitative linguistic features from children's spoken language and trains machine learning classifiers to distinguish ASD cases.

To better capture cross-modal interactions, Tsai et al. [60] proposed a multimodal fusion framework (MULT) that integrates visual, audio, and textual features via cross-modal attention mechanisms. Although these approaches generally outperform unimodal baselines, they still follow a feature extraction–classification pipeline that directly maps representations to diagnostic labels. Such pipelines lack explicit modeling of structured clinical rules like ADOS and often cannot simultaneously predict multiple item-level scores, thereby limiting interpretability and clinical applicability. Moreover, the lack of in-domain training data often further degrades the performance.

### 2.2. Fine-grained behavioral modeling for ADOS-related assessment

Unlike direct diagnostic classification, ADOS evaluates multiple dimensions of social communication and interaction through structured paradigms and standardized item-level scoring rules. Consequently, recent studies have attempted to move beyond coarse diagnostic prediction toward modeling the fine-grained behavioral patterns associated with ADOS assessments.

Early research primarily focused on identifying and evaluating specific interactive behaviors closely related to ADOS assessment items. For instance, Pan et al. [52] developed a multimodal machine learning framework that integrates speech recognition and facial expression analysis to automatically evaluate children's responses in the Responsive Social Smiling paradigm. Zhang et al. [73] constructed a 3D spatiotemporal facial analysis pipeline with few-shot learning, extracting facial dynamics from ADOS interview videos through spatiotemporal feature and scene-level fusion. Cheng et al. [10] proposed a computer-assisted ASD detection system based on multimodal behavioral signal processing, in which audio-visual signals are parsed to evaluate several ADOS-related behavioral patterns. Kim et al. [26] introduced an AI-based screening framework that analyzes structured home videos (e.g., name response, imitation, and ball play) by extracting gaze, motion, and audio features to detect ASD risk. Kinfu et al. [27] proposed CAMI-2DNet, a deep learning framework that assesses motor imitation deficits by analyzing motion representations from video-based imitation tasks. Although these studies attempt to model clinically relevant behavioral cues rather than perform direct diagnostic classification, they typically focus on a limited subset of behaviors and therefore cannot provide a comprehensive assessment across the full range of ADOS-related dimensions.

More recently, script-centric behavior understanding methods [37] have emerged as a promising pipeline. This approach converts audio-visual data into structured behavioral scripts, enabling LLMs to perform downstream reasoning. By bridging multimodal perception models and LLMs through textual behavioral representations, script-based frameworks offer improved interpretability and better alignment with clinical assessment rules. However, this method is primarily used to map multimodal inputs to behavior labels or textual descriptions, without explicitly modeling standardized clinical assessment protocols or leveraging structured dialogue-level semantics.

### 2.3. Multimodal large language models for clinical reasoning

LLMs have recently achieved remarkable progress in instruction following, reasoning, and domain generalization. Their strong natural language processing capabilities demonstrate considerable potential for clinical reasoning [1,64,70]. In addition to text-only LLMs, Multimodal Large Language Models (MLLMs) have emerged that can jointly process visual, auditory, and textual inputs. For example, Zhang et al. [72] proposed Video-LLaMA, an instruction-tuned audiovisual language model that aligns pretrained visual and audio encoders with a large language model through a Q-former module, enabling unified understanding of multimodal video content. Li et al. [29] introduced a biomedical vision–language model that jointly trains a CLIP visual encoder with a large language model via instruction fine-tuning, enabling conversational understanding and visual question answering on medical images. Jiang et al. [25] proposed Med-MoE, a lightweight medical vision–language framework that incorporates a hybrid expert architecture and a modality-aware routing mechanism to support efficient multimodal medical reasoning across tasks. Deng et al. [12] introduced an audio–visual autism behavior recognition framework that combines visual features, audio captions, and speech transcripts with a multimodal large language model to detect autism-related behaviors and generate interpretable explanations. He et al. [22] developed PFMVG, a parameter-efficient fine-tuning framework for medical multimodal LLMs. Built upon LoRA, this framework adapts LLMs to visual localization tasks, enabling precise alignment between medical textual descriptions and lesion regions in medical images.

Despite these advances, directly applying multimodal LLMs to structured behavioral assessment remains challenging. First, existing approaches do not explicitly model standardized clinical assessment procedures such as ADOS. In contrast, our work decomposes the problem into paradigm classification, item-level behavioral assessment, and diagnostic inference, which closely align with real-world clinical evaluation protocols. Second, prior studies mainly rely on low-level audio features or plain speech transcripts, lacking explicit modeling of dialogue-level semantic information, which is critical for assessing social communication behaviors.

### 2.4. Multimodal adaptation and spatiotemporal modeling

Recent advances in multimodal learning have substantially improved the capabilities of multimodal large language models through parameter-efficient adaptation techniques and more effective spatiotemporal representation learning. For multimodal adaptation, Shen et al. [55] propose a conditional mixture of LoRA experts to mitigate task interference by constructing input-specific low-rank adaptations. Wang et al. [61] introduce multimodal prompt tuning that learns prompts on both the vision and language sides to improve cross-modal alignment. Liu et al. [40] formulate efficient multimodal instruction tuning as directly editing semantically rich multimodal representations. Recent video understanding research has further explored spatiotemporal modeling and parameter-efficient reasoning. Yan et al. [67] develop a weakly supervised video instance segmentation framework that leverages spatiotemporal collaboration. Yan et al. [66] propose knowledge-aware generative prompt collaboration for parameter-efficient video captioning. Chen et al. [7] introduce a prompt-guided intent aligner for lightweight iterative reasoning in zero-shot object navigation. Yan et al. [68] integrate long-range, short-range, and local-keyframe representations for fine-grained video captioning. Although these methods address tasks different from clinical behavioral assessment, they highlight promising directions for modeling temporal dynamics, adapting multimodal representations, and improving structured reasoning over video inputs.

## 3. ADOS-M1 clinical dataset collection

The SYSU-ADOS-M1 clinical dataset is a multi-view audio-visual dataset designed in accordance with the ADOS Module 1 [43] for structured ASD assessment. All participants provided written informed consent, which was approved by the Institutional Review Board (IRB) of the Third Affiliated Hospital of Sun Yat-sen University.

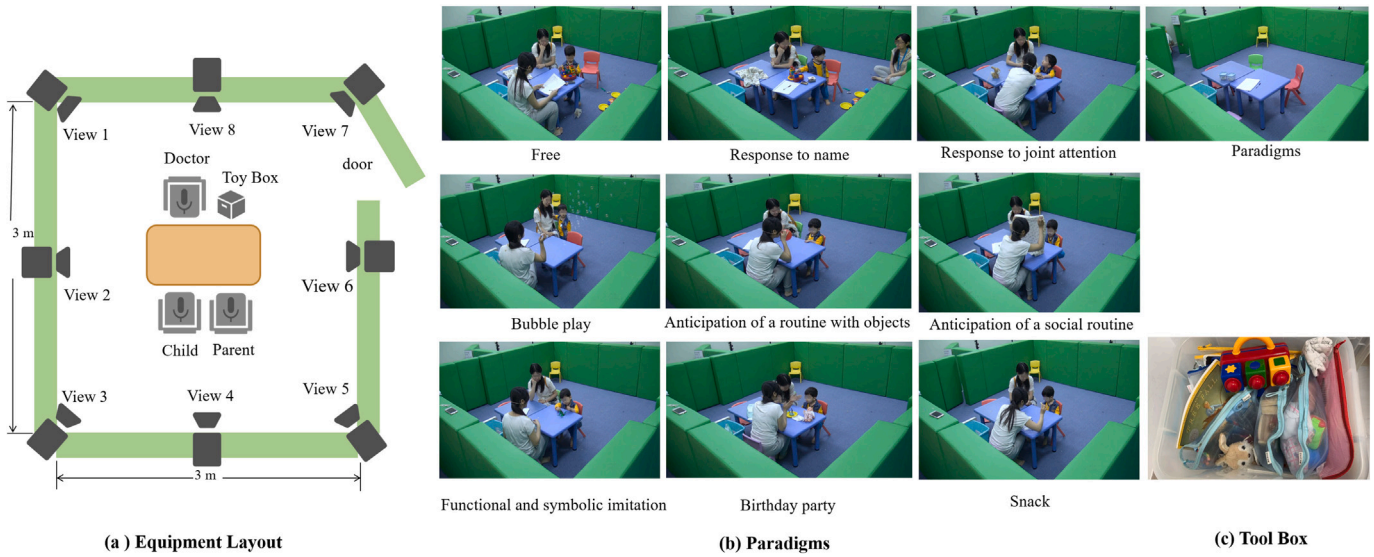



**Fig. 1.** (a) Equipment layout of the experimental scene. The experimental site was enclosed within a 3 m × 3 m × 1 m space using some crash fences to ensure uniform data collection. Eight cameras were evenly distributed around the area to enable unobstructed capture of the child's behavior during the ADOS assessment. The table and chairs in the center were provided for the doctor, child, and parents to complete the ADOS assessment. A toy box was also included to store toys that might be used during the assessment. Additionally, each participant wore a microphone to collect voice signals clearly. (b) Snapshots of the ten paradigms captured from Viewpoint 1 as shown in (a). Although multiple views were collected, only Viewpoint 1 was used to ensure consistency and reduce computational complexity. (c) Overview of the toys used in the assessment paradigms.

### 3.1. Capture system

To minimize external influences on the ADOS assessment process and ensure video stability, all recordings were conducted in a quiet hospital room with a standardized data acquisition system. Fig. 1 (a) illustrates the equipment setup. Eight RGB cameras (1280 × 768 resolution) were placed around the assessment area, capturing synchronized 50 fps recordings. In post-processing, video streams were further aligned using recorded timestamps to ensure temporal consistency across views.

The doctor and child sat facing each other at a table, with the parent seated beside the child for support. The doctor and child each wore wireless microphones, capturing individual speech signals at a 16 kHz sampling rate, resulting in three synchronized audio tracks. These audio tracks were subsequently merged into a single audio stream for downstream processing.

### 3.2. Paradigms

Module 1 of the ADOS [43] is designed for toddlers or children with very limited language skills. The doctor employs a series of semi-structured activities to observe the child's socialization, communication, imaginative play, and imitation skills. The ADOS-M1 video dataset comprises 10 structured interaction sections (paradigms) between doctors and children. Fig. 1 (b) illustrates the snapshots of the ten paradigms. The experimental paradigms and their corresponding descriptions are shown in Table 1.

**Table 1**
Description of Experimental Paradigms.

| Paradigm | Abbreviation | Description |
|---|---|---|
| Free | FR | The child engages in spontaneous play with age-appropriate toys in an unstructured setting, enabling assessment of natural social initiation toward the parent or doctor. |
| Response to Name | RN | The child's attention and responsiveness to external social cues are evaluated by calling their name during ongoing activity. |
| Response to Joint Attention | RJA | Shared attention is elicited via gaze, gestures, or verbal cues, assessing the child's ability to follow and coordinate attention with the doctor. |
| Bubble Play | BUP | Interactive engagement is stimulated through bubble play, capturing affective expression, attention, shared enjoyment, and motor responses. |
| Anticipation of a Routine with Objects | ARO | The child's ability to anticipate sequential actions is assessed within familiar object-based routines (e.g., balloon play). |
| Anticipation of a Social Routine | ASR | Anticipatory behavior is evaluated within familiar social interactions (e.g., hide-and-seek), reflecting social prediction ability. |
| Functional and Symbolic Imitation | FSI | The child's capacity to imitate familiar functional and symbolic actions demonstrated by the doctor is examined. |
| Birthday Party | BIP | A simulated social scenario is used to assess social interaction, emotional expression, and verbal communication abilities. |
| Snack | SN | Food-related interaction is used to evaluate the child's requesting behavior and responsiveness to others. |
| Non-paradigm | NP | Unstructured segments outside predefined paradigms, typically corresponding to transitions between adjacent activities. |

### 3.3. ADOS participants

During the recruitment phase, a total of 628 parents agreed to have their children participate in this study. During the data screening phase, we excluded children diagnosed with Fragile X syndrome, tuberous sclerosis, Down syndrome, or those with a significant early medical history. Additionally, children with missing data recordings or incomplete basic information were excluded. The final dataset included 447 participants. It comprises 316 children diagnosed with ASD and 131 children with TD. Participants ranged in age from 17 to 161 months, including 366 males and 81 females. Although a small number of participants exceeded the typical age range of Module 1, their expressive language abilities were consistent with Module 1 criteria. The mean age was 39.9 months for male participants and 44.6 months for female participants. Fig. 2 clearly shows that ages are concentrated between 24 and 60 months.

All participants were Mandarin-speaking or Cantonese-speaking children from an urban region of China. Although specific socioeconomic

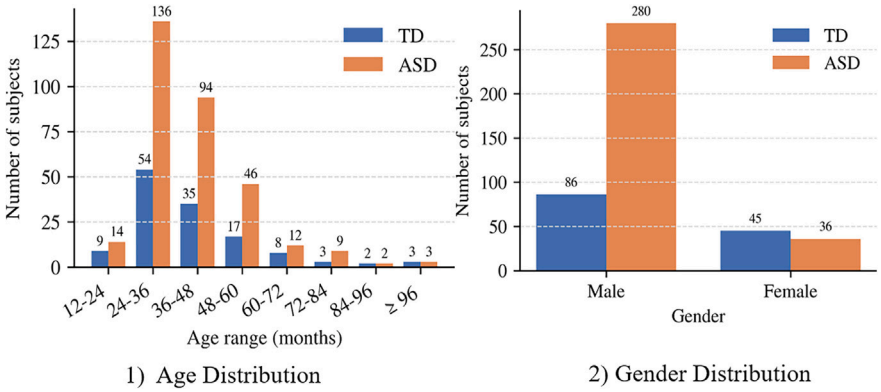



**Fig. 2.** Statistics of ASD/TD groups in the clinical database. Barplot graphs demonstrate the age and gender distributions of children with ASD and TD.

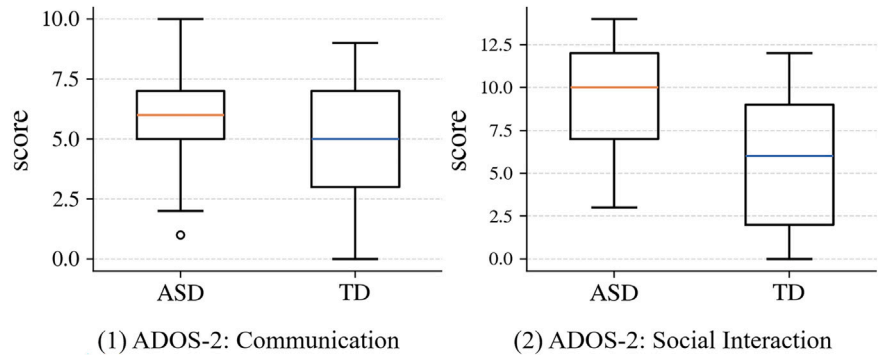



**Fig. 3.** The Boxplots of (a) communication and (b) social scores for TD and ASD groups. Boxes indicate the interquartile range, and the horizontal line denotes the median.

indicators were not stored due to privacy regulations, the recruitment center predominantly serves middle-income families.

Each child underwent a standardized ADOS-2 Module 1 assessment [20]. ADOS-2 assessments were administered by experienced doctors who had obtained clinical administration qualifications for ADOS-2. Final ASD diagnostic labels were determined through consensus among at least two senior doctors based on a comprehensive evaluation of ADOS-2 scores, clinical observations, and developmental history. Fig. 3 displays the scoring distribution of the ADOS-2 for child participants, providing a basic profile of children in our clinical database.

### 3.4. ADOS scores and annotation

The expert scores for ADOS Module 1 are determined by evaluating all paradigm videos collectively, rather than assessing individual paradigm videos. The scoring system consists of 29 items categorized into 5 categories: (A) language and communication (8 items), (B) reciprocal social interactions (12 items), (C) play (2 items), (D) stereotypic behaviors and restricted interests (4 items), and (E) other unusual behaviors (3 items). In this study, only 12 items (A2, A5, A6, A7, A8, B1, B3, B5, B9, B10, B11, B12) relevant to rule-based ASD classification algorithm [20] were selected. The specific item names and their corresponding descriptions are provided in Table 14 (Appendix A).

Each item may contain several score levels: 0, 1, 2, 3, 7, and 8. Most items use scores ranging from 0 to 3, representing the severity of ASD-related behaviors. A score of 0 indicates that the behavior does not align with the expected ASD characteristics for that item, whereas a score of 3 signifies a high level of autism-related concern. A small subset of items includes scores of 7 or 8, which denote that the behavior is not applicable for ASD assessment. For example, a score of 8 for item A3 indicates insufficient vocalization, preventing speech evaluation, while a score of 7 for item E1 suggests an activity level too low for judgment.

According to the revised ADOS-2 assessment algorithms [20], item-level assessment scores (item scores) are recoded as follows: (1) item scores of 0, 1, and 2 are retained, (2) item scores of 3 are recoded as 2, (3) item scores of 7 and 8 are recoded as 0. After recoding, the revised algorithm is used to compute domain totals for Language and Communication (Class A) and Reciprocal Social Interaction (Class B). Diagnostic classifications are then determined based on established cutoff thresholds using Class A, Class B, and their combined total (A + B) [51]. Algorithm 1 illustrates the process of score recoding and diagnostic classification in Module 1. This assessment algorithm is applied solely for diagnostic classification.

### 3.5. ADOS assessment videos

The ADOS is a semi-structured, standardized behavioral observation scale. Trained doctors systematically elicit and observe an individual's communication, social interaction, and repetitive behaviors across different interaction paradigms. In this study, a standardized data collection system recorded video and audio of the entire interaction. Most of the time, participants sit in chairs with the doctor and conduct the interaction at a table, ensuring the camera can clearly capture them. Each video consists of 20–80 min. The average duration of the videos is about 36.38 min.

Due to computational resource constraints, long videos remain challenging for direct model analysis. Therefore, the long videos were divided into non-overlapping 30-second clips. Based on the annotated start and end timestamps for each paradigm, the type of paradigm of each 30-second clip can be derived from the manually labeled timestamps. Table 2 presents the average length, total length, and the number of 30-second clips for each paradigm.

Because the raw audiovisual recordings contain sensitive information involving children, they cannot be released publicly. Subject

**Algorithm 1** Calculation and Classification for ADOS-M1 [51].

**Input:** Category A scores $A2, A5, A6, A7, A8$; Category B scores $B1, B3, B5, B9, B10, B11, B12$

**Output:** Diagnostic classification $d \in \{\text{Autism}, \text{Autism Spectrum}, \text{Non-Spectrum}\}$

$S_A \leftarrow A2 + A5 + A6 + A7 + A8$
$S_B \leftarrow B1 + B3 + B5 + B9 + B10 + B11 + B12$
**if** $S_A \geq 4 \wedge S_B \geq 7 \wedge S_A + S_B \geq 12$ **then**
  $d \leftarrow$ Autism
**else if** $S_A \geq 2 \wedge S_B \geq 4 \wedge S_A + S_B \geq 7 \wedge \neg(S_A \geq 4 \wedge S_B \geq 7 \wedge S_A + S_B \geq 12)$ **then**
  $d \leftarrow$ Autism Spectrum
**else**
  $d \leftarrow$ Non-Spectrum
**end if**
**return** $d$

**Table 2**
The information of selected ADOS paradigm videos.

| Paradigm | Average Length | Total Length | Clip Count |
|---|---|---|---|
| FR | 9.79 | 4484.90 | 7385 |
| RN | 0.42 | 660.13 | 510 |
| RJA | 1.28 | 601.56 | 663 |
| BUP | 2.35 | 1112.45 | 1765 |
| ARO | 2.56 | 1245.08 | 2502 |
| ASR | 0.82 | 585.06 | 569 |
| FSI | 1.94 | 935.52 | 1374 |
| BIP | 5.19 | 2296.17 | 4053 |
| SN | 4.04 | 1895.05 | 3325 |
| NP | 5.29 | 2366.9 | 8017 |

1 Average length is computed per subject.
2 Total length is aggregated across all participants.
3 The statistical units for "Average Length" and "Total Length" are minutes.

to institutional and ethical approval, we will work toward providing qualified research collaborators with controlled access to de-identified behavioral feature representations and corresponding ADOS item-level scores.

### 3.6. Task design

Clinical assessment videos for a single individual are typically up to one hour long and contain both paradigm and non-paradigm clips. To assess long videos fully automatically, we divide them into non-overlapping 30-second clips. In the first step, each clip is classified into a paradigm category to filter out non-paradigm segments and to provide category information for the subsequent assessment model. In the second step, multi-item assessments are conducted on the paradigm clips. Therefore, the task design is as follows:

**Task 1: Paradigm Classification.** The first task aims to identify the interaction paradigm of a given 30-second video clip. Each clip is assigned a single paradigm label corresponding to the ADOS-2 Module 1 activity during which it was recorded. This task evaluates the model's ability to distinguish between structured and unstructured assessment paradigms.

**Task 2: Item-level Behavioral Assessment.** The second task focuses on predicting assessment scores for paradigm clips (Non-paradigm clips are excluded in advance during both training and testing). For paradigm clips, the model predicts scores for all ADOS behavioral items. Each item score is formulated as a multi-class classification problem, where discrete score values follow the standardized ADOS-2 scoring criteria. Although ADOS item-level scores are determined based on the entire assessment session rather than individual video segments, we apply the same item scores to all 30-second clips extracted from the same recording. This design is motivated by the observation that many ADOS behavioral indicators, such as vocalization frequency, joint attention, and social interaction patterns, tend to manifest repeatedly across different paradigms during the assessment process. As a result, short video clips still contain informative behavioral cues relevant to the final item-level evaluation. Moreover, using clip-level supervision enables the model to learn fine-grained behavioral patterns associated with different interaction paradigms while preserving standardized clinical labels. During inference, predictions from multiple clips are aggregated to obtain subject-level behavioral assessments.

In Task 2, original ADOS item scores, including values of 7 and 8, are treated as distinct classes and directly learned by the model in a multi-class classification setting. The recoding of assessment scores 7 or 8 to 0 is applied only during the post-processing stage for ASD diagnostic classification. All models predict original item scores rather than recoded scores.

The two tasks are designed to capture complementary aspects of clinical assessment. Paradigm classification provides coarse-grained contextual understanding, while behavioral score prediction targets fine-grained behavioral quantification. The integration of these two tasks enables fully automated assessment of long videos.

## 4. Methods

Fig. 4 illustrates the general framework of our proposed multimodal method, enabling LLMs to recognize ADOS paradigm and assess ADOS behavior. This method enables LLMs to understand visual and audio signals to accomplish specific ADOS-related tasks.

Specifically, we designed three branches: the visual branch, audio branch, and dialogue branch, to convert video and audio into token representations comprehensible to LLMs. In this section, we first define the problem, then introduce the structure and functionality of each branch, and finally describe large language model and the process of instruction tuning.

### 4.1. Problem formulation

Each sample corresponds to a 30-second multimodal clip $x$, including a video segment $v$, and an audio segment $a$. We study two tasks based on the same multimodal input:

- **Task 1 (Paradigm Classification):** Given a 30-second clip, predict its corresponding ADOS paradigm label $y^{(p)} \in \{1, \ldots, K\}$, where $K$ includes all structured paradigms as well as the non-paradigm (NP) class.
- **Task 2 (Item-level Behavioral Assessment):** Given a 30-second clip from paradigm clips, for each ADOS behavioral item $m \in \{1, \ldots, M\}$, predict a discrete score $y_m^{(i)} \in \{0, 1, 2, 3, 7, 8\}$.

Task 1 can be formulated as a multi-class classification problem with 10 categories, while Task 2 can be regarded as a multi-task multi-class classification problem involving 12 ADOS items, each associated with 4 or 5 possible score levels. Both tasks were reformulated as structured text generation problems for LLMs. The generated structured outputs were then parsed using regular expressions to obtain independent prediction results.

### 4.2. Visual branch

The video branch takes a 30-second video clip as input, from which visual representations are extracted and subsequently analyzed by the backend LLM. When selecting a video foundation model, it is essential to consider its capability to model temporal dynamics and support complex video–language understanding tasks. To this end, we adopt the recently proposed video foundation model (UML [32]) for visual feature extraction. UML is built upon the vanilla ViT-L/16 [14] architecture and employs a progressive pre-training framework. Benefiting from large-scale pre-training on the Kinetics-710 dataset [31] and multiple image–text corpora [8], UML is well suited for capturing temporal information and supporting time-aware video–language understanding.

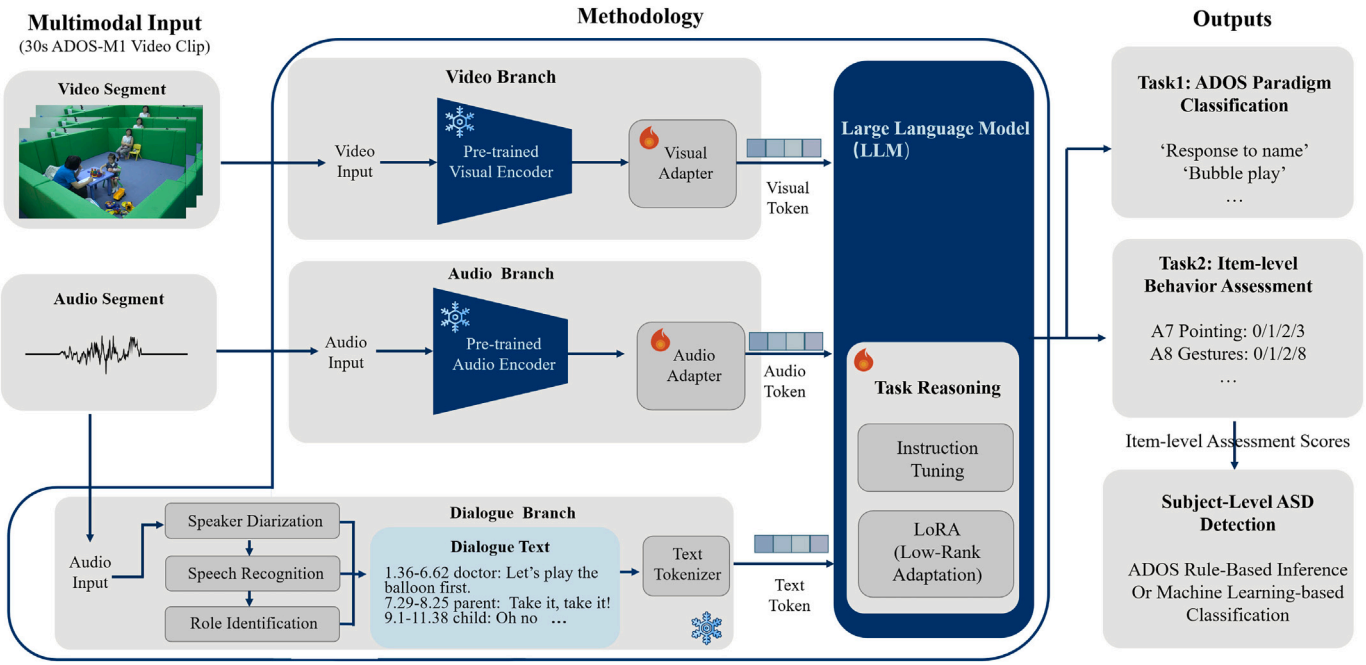



**Fig. 4.** The overview of our proposed Visual-Audio-Dialogue (VAD) framework for ADOS-M1 behavioral assessment. Given a 30-second ADOS-M1 video segment, the framework integrates three modalities, including visual, audio, and dialogue information. The visual and audio streams are encoded by pre-trained encoders and aligned to the language space via modality-specific adapters. In parallel, structured dialogue text is obtained through speaker diarization, speech recognition, and role identification, followed by tokenization. All modality tokens are then fed into a large language model (LLM), where instruction tuning and LoRA are employed to enhance task-specific reasoning. The model jointly performs two tasks: ADOS paradigm classification and item-level behavioral assessment. The predicted item scores are further utilized for subject-level ASD detection via rule-based inference or machine learning-based classification.

Given a video segment $v$, we first downsample the frame rate to 8 FPS and then extract visual tokens using a UML encoder, which is kept frozen during training. The resulting visual token sequence is denoted as

$$\mathbf{Z}_v \in \mathbb{R}^{T_v \times d_v}, \tag{1}$$

where $T_v$ and $d_v$ denote the number of tokens and the embedding dimension, respectively. In our setting, $T_v = 240$ and $d_v = 1024$.

Pre-trained visual encoders provide powerful general representations. However, their feature distributions differ significantly from those of LLMs. To address this, we introduce lightweight visual-modal adapters that project visual representations onto a unified semantic space grounded in language reasoning, enabling efficient cross-modal fusion.

The visual adapter is implemented as a lightweight linear projection layer that maps UML features to the target embedding dimension of the backend language model. This design ensures minimal parameter overhead while enabling effective alignment between visual tokens and language representations. The lightweight linear projection is applied to align visual tokens:

$$\mathbf{H}_v = f_v(\mathbf{Z}_v) = \mathbf{Z}_v \mathbf{W}_v, \tag{2}$$

where $\mathbf{W}_v \in \mathbb{R}^{d_v \times d}$ is a trainable projection matrix, and $\mathbf{H}_v \in \mathbb{R}^{T_v \times d}$.

### 4.3. Audio branch

The audio branch takes the audio corresponding to a 30-second video clip as input, extracts an audio representation from it, and then feeds it to the subsequent LLMs. Verbal and nonverbal cues play a crucial role in behavioral assessment, particularly in tasks involving social communication and interaction analysis. In clinical assessment scenarios, audio signals often exhibit high variability in speech rate, articulation, background noise, and recording conditions. Therefore, a pretraining audio encoder with strong robustness and generalization capabilities is essential for reliable feature extraction. In this work, we adopt the Whisper-large-v3 [53] model as the audio backbone to extract audio features from raw speech signals.

Given an audio segment $a$, we extract audio features using Whisper-large-v3, which is also frozen during training:

$$\mathbf{Z}_a \in \mathbb{R}^{T_a \times d_a}. \tag{3}$$

For audio adapter, we adopt the *OsumAdaptorModule* proposed in OSUM [17], which consists of 3-layer 1D convolutional layers (Conv1D) and 4-layer Transformer, and a linear projection layer to map audio features into the LLM embedding space:

$$\mathbf{H}_a = f_a(\mathbf{Z}_a), \quad \mathbf{H}_a \in \mathbb{R}^{T_a \times d}. \tag{4}$$

This architecture bridges the output of the speech encoder with the input requirements of the LLM, achieving efficient modal alignment.

### 4.4. Dialogue branch

The dialogue branch takes a raw audio signal as input and aims to transform audio into structured dialogue through a multi-stage speech processing pipeline. Specifically, the speech signal is processed in three sequential stages.

In the first stage, a speaker diarization model, S2SND [9], is employed to segment the continuous audio stream and generate a timestamped speech log $d$, where speaker role identities are initially unknown. In the second stage, the automatic speech recognition model (WhisperX [4]) is applied to the speech segments defined by the speech log, producing timestamped dialogue transcripts $s$. In the third stage, a speech-based age estimation model (wav2vec2 [6])is used to infer the age of each speech segment. Given that the video recordings involve only three predefined role identities (doctor, parent, and child) and follow a semi-structured evaluation protocol, speaker roles are inferred

using a rule-based strategy that combines age estimation and speaking frequency statistics. Specifically, doctors are typically characterized by higher speaking frequency and older age, parents by lower speaking frequency and older age, and children by younger age.

Finally, the dialogue content is organized into a structured format [timestamp, speaker, utterance], as illustrated in the Dialogue Text of Fig. 4. The dialogue content is tokenized using the LLM tokenizer to produce text tokens $\mathbf{Z}_t$. Similarly, specific task prompts and questions are also processed by the LLM tokenizer and converted into text tokens, referred to as $\mathbf{P}_t$ and $\mathbf{Q}_t$, respectively.

### 4.5. Large language models

Finally, we concatenate the adjusted visual token $\mathbf{H}_v$, audio token $\mathbf{H}_a$, and text token $\mathbf{H}_t$ (including task-specific prompts $\mathbf{P}_t$, questions $\mathbf{Q}_t$, and dialogues $\mathbf{Z}_t$) to form the input sequence for LLMs:

$$\mathbf{X} = [\mathbf{H}_v, \mathbf{H}_a, \mathbf{P}_t, \mathbf{Q}_t, \mathbf{Z}_t] \tag{5}$$

The objective is to generate reasonable and consistent target outputs (answers) $\mathbf{X_a}$ based on $\mathbf{X}$.

$$\mathbf{X_a} = LLM(\mathbf{X}) \tag{6}$$

We employ LLMs based on the decoder-only architecture (such as the Qwen series models [69]) to process input sequences and generate structured text responses. For Task 1, the model outputs a single paradigm label; for Task 2, it outputs assessment scores for the queried ADOS items.

### 4.6. Multimodal instruction tuning

The multimodal instruction fine-tuning in this paper comprises two stages: In the first stage, the parameters of the visual encoder, audio encoder, and LLM (all checkpoints from existing open-source models) are frozen, while only the parameters of the video and audio adapters are trained to achieve multimodal language alignment [12,33,72]. In the second stage, the frozen LLM parameters are unfrozen, and LoRA [24] is employed for efficient training to adapt to different downstream tasks [35]. In both stages, we fine-tune the model using instruction data to optimize performance. To adapt to different tasks, the instruction content is designed differently, with specific instructions detailed in Appendix B. We perform instruction-tuning of the LLM on the prediction tokens, using its original auto-regressive training objective. The training process is given a target text answer $y$, and the training objective is the standard next-token prediction loss:

$$\mathcal{L} = -\sum_{t=1}^{|y|} \log p_\theta(y_t \mid \mathbf{X}, y_{<t}). \tag{7}$$

Fig. 5 illustrates the unified format for multimodal instruction sequences during multi-turn dialogues, showing only two dialogue turns here. The number of turns varies by task category: Task 1 has only one turn, while Task 2 has 12 turns, each corresponding to one ADOS item assessment. During training, we follow the format specified by the Qwen model [69] to structure the instruction sequence. After fine-tuning, the LLM predicts both the assistant response and the stop position; thus, the green sections in the format are used to compute losses in the auto-regressive model.

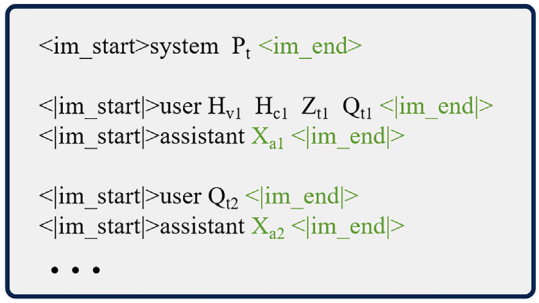



**Fig. 5.** The unified format for multimodal instruction sequences. `<|im_start|>system` denotes the start marker for system prompts, `<|im_start|>user` indicates the start marker for user utterances, and `<|im_start|>assistant` signifies the start marker for model responses. `<|im_end|>` denotes the stop token.

## 5. Experiments

### 5.1. Experimental setup

All experiments adopt a subject-independent data splitting strategy. Specifically, participants are randomly sampled at the subject level into training and test sets with a ratio of 90% and 10%, respectively. All 30-second video clips extracted from the same participant are strictly assigned to the same split, ensuring that no subject appears in both training and test sets.

A fixed random seed is used to ensure reproducibility, and the main experiments use the same split to maintain a consistent evaluation protocol across tasks. To further assess robustness across alternative subject splits, we additionally perform subject-independent stratified five-fold cross-validation, as reported in Section 5.5.5.

### 5.2. Baseline methods

We established a comprehensive benchmark comparison, as shown in Table 1. The baseline methods encompass two categories: feature-based deep model methods and script-centric behavior understanding (SCBU) methods [37]. The specific experimental workflow and details are as follows:

(1) **Feature-based deep learning methods.** This class of methods follows an end-to-end feature classification framework across multiple modalities. Modality-specific representations are first extracted using deep learning models, after which prediction heads are employed to directly learn the mapping from input features to target labels for various downstream tasks. In our experiments, unimodal approaches (visual, audio, and dialogue) as well as tri-modal fusion methods were evaluated separately. To ensure a fair and rigorous comparison, the pre-trained models used for feature extraction were kept consistent with those adopted in our proposed framework.

- Vision Transformer (ViT) [32] is employed to extract visual representations, followed by training a prediction head for paradigm classification or ADOS item score prediction. The extracted video features have a dimensionality of $1568 \times 1024$ per second. For a 30-second clip, frame-level features are temporally aggregated via average pooling to produce a sequence-level representation.
- Whisper-Large-V3 [53] is employed to extract audio representations, followed by training a prediction head for paradigm classification or item-level assessment score prediction. The extracted audio features have a dimensionality of $1500 \times 1280$ per second. Temporal aggregation is performed using average pooling along the time dimension to obtain a compact clip-level representation.
- BERT [13] is employed to extract text representations, followed by training a prediction head for paradigm classification or ADOS item score prediction. Specifically, the textual modality is derived from timestamped dialogue content obtained through the dialogue branch of the proposed framework. For each 30-second audio segment, transcripts are temporally aligned by dividing the segment into fixed-length intervals (1 s), with speech segments assigned to corresponding intervals based on timestamps. Intervals without dialogue are filled with a special token ($\langle silence \rangle$) to preserve temporal consistency. Each interval-level text is then encoded using a pre-trained BERT model, from which the [CLS] token representation is extracted as the segment-level embedding. This process produces a fixed-length textual feature of size $30 \times 768$.

**Table 3**
The accuracy for paradigm classification (10-category) at the 30-second clip level in Task 1. V represents video, A represents audio, T represents text, and D represents dialogue transcripts.

| Model | Modality | Post-training | Acc |
|---|---|---|---|
| VIT [32] | V | ✓ | 0.2629 |
| Whisper-Large-V3 [53] | A | ✓ | 0.4505 |
| BERT [13] | T | ✓ | 0.4099 |
| MULT [60] | VAT | ✓ | 0.5573 |
| SCBU-qwen3-omni [37] | VAT | × | 0.4763 |
| VAD-Qwen2.5-0.5B | VAD | × | 0.3945 |
| | | ✓ | 0.6091 |
| VAD-Qwen3-0.6B | VAD | × | 0.4195 |
| | | ✓ | 0.6119 |
| VAD-Qwen3-4B | VAD | × | 0.4498 |
| | | ✓ | 0.6588 |
| VAD-Qwen3-8B | VAD | × | 0.4846 |
| | | ✓ | 0.6763 |
| VAD-Qwen3-30B | VAD | × | 0.4994 |
| | | ✓ | 0.7377 |

- The multimodal transformer (MULT) [60] is employed for temporal cross-modal fusion of multimodal features, followed by training a prediction head for paradigm classification or ADOS item score prediction. To address the mismatch in feature dimensions across modalities, we apply a learnable linear projection layer prior to cross-modal fusion.

(2) **Script-centric behavior understanding method.** This method leverages a general video-language model to transform multimodal inputs into structured behavioral scripts, thereby enabling large language models (LLMs) to perform downstream tasks. Specifically, we employ Qwen-Omni [65] to convert audio and video data into textual representations, which are then fed into Qwen3-30B-A3B-Instruct [69] to accomplish Task 1 and Task 2.

### 5.3. Overall performance

We evaluate the proposed Visual-Audio-Dialogue (VAD) framework on both Task 1 (Paradigm Classification) and Task 2 (Item-level Behavioral Assessment). Within the VAD framework, we evaluated the impact of LLMs with different parameter scales. For example, VAD-Qwen2.5-0.5B denotes that the LLM backbone adopts the Qwen2.5 model with 0.5B parameters. The quantitative results are reported in Tables 3 and 4, respectively.

**Task 1: Paradigm Classification.** As shown in Table 3, unimodal models exhibit limited performance. The visual-only model (ViT) achieves an accuracy of 26.29%, while audio-only (Whisper-Large-V3) and text-only (BERT) models obtain 45.05% and 40.99%, respectively. This indicates that visual cues alone are insufficient for reliable paradigm identification, and speech-related modalities provide more discriminative information.

Among traditional multimodal baselines, MULT achieves 55.73% accuracy, outperforming the zero-shot SCBU-Qwen3-Omni. A potential explanation for the inferior performance of zero-shot SCBU compared with traditional multimodal feature-based methods is its reliance on the capabilities of the video-language models. Despite recent advances, current video-language models remain limited in their ability to directly capture fine-grained distinctions between ASD-related behaviors and different ADOS paradigms.

However, when integrating multimodal features with instruction tuning and LoRA, the performance improves substantially. In particular, VAD-Qwen3-30B with LoRA achieves the best accuracy of 73.77%, surpassing the strongest baseline (MULT) by 18.04% in absolute terms. This significant improvement demonstrates the effectiveness of multimodal alignment and large-scale language reasoning for 10-category paradigm classification.

For the impact of LoRA Fine-Tuning, the accuracy of LLMs across different parameter scales improves significantly after LoRA fine-tuning. This substantial and stable improvement suggests that LoRA fine-tuning plays a crucial role in adapting LLMs for paradigm classification tasks. Without post-training, even large-scale models fail to fully exploit multimodal cues. For example, VAD-Qwen3-30B achieves 47.63%, which is lower than MULT. With LoRA fine-tuning, the models can effectively align visual-audio-text representations with the target paradigm categories.

Furthermore, model scaling remains beneficial under LoRA fine-tuning. The accuracy steadily increases from 60.91% (VAD-Qwen2.5-0.5B) to 73.77% (VAD-Qwen3-30B), demonstrating that both LoRA and model capacity contribute to improved paradigm understanding.

**Task 2: Item-level Behavioral Assessment.** Table 4 reports the accuracy for each ADOS item and the overall mean accuracy. Compared with unimodal and conventional multimodal baselines, our method achieves substantially better performance across almost all items. The best-performing model, VAD-Qwen3-30B with LoRA, achieves a mean accuracy of 72.61% at the 30-second clip level, which is 25.60% higher than the strongest baseline BERT (47.01%).

Notably, performance gains are consistently observed across both socially interactive items (A-series) and communicative behavior items (B-series). For the A-series, which typically involve language and nonverbal interaction cues, the proposed model yields substantial improvements. For example, the accuracy on A2 increases from 24.22% (VIT) to 93.89% (VAD-Qwen3-30B with LoRA), and A8 improves from 26.41% (BERT) to

**Table 4**
The accuracy of each ADOS item at the 30-second clip level in Task 2. Language and communication items are denoted in A-series and reciprocal social interaction are B-series items (Details can be found in Table 14). Post represents Post-training.

| Model | Post | A2 | A5 | A6 | A7 | A8 | B1 | B3 | B5 | B9 | B10 | B11 | B12 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VIT [32] | ✓ | 0.2422 | 0.4513 | 0.4238 | 0.2396 | 0.3206 | 0.7899 | 0.3929 | 0.4882 | 0.4577 | 0.4814 | 0.3865 | 0.2441 | 0.4098 |
| Whisper-Large-V3 [53] | – | 0.3458 | 0.4950 | 0.4475 | 0.3744 | 0.3228 | 0.8140 | 0.4867 | 0.4188 | 0.5228 | 0.5801 | 0.4520 | 0.3345 | 0.4662 |
| BERT [13] | ✓ | 0.2475 | 0.5763 | 0.5714 | 0.3850 | 0.2641 | 0.8535 | 0.2968 | 0.4241 | 0.5624 | 0.6584 | 0.4776 | 0.3251 | 0.4701 |
| MULT [60] | ✓ | 0.2870 | 0.3378 | 0.5171 | 0.2934 | 0.2915 | 0.8026 | 0.4915 | 0.3548 | 0.4667 | 0.5242 | 0.4606 | 0.3559 | 0.4319 |
| SCBU-Qwen3-Omni [37] | × | 0.3543 | 0.3974 | 0.2294 | 0.2983 | 0.4148 | 0.8338 | 0.5635 | 0.1363 | 0.5869 | 0.6434 | 0.1174 | 0.3501 | 0.4105 |
| VAD-Qwen2.5-0.5B | × | 0.3185 | 0.4048 | 0.4727 | 0.3061 | 0.4167 | 0.7641 | 0.4727 | 0.4727 | 0.4690 | 0.4543 | 0.4130 | 0.2331 | 0.4331 |
| | ✓ | 0.5286 | 0.5667 | 0.6135 | 0.5318 | 0.5557 | 0.8054 | 0.6406 | 0.5699 | 0.5814 | 0.5594 | 0.5796 | 0.5291 | 0.5885 |
| VAD-Qwen3-0.6B | × | 0.3070 | 0.4094 | 0.4484 | 0.3199 | 0.3974 | 0.7737 | 0.4263 | 0.4897 | 0.4924 | 0.4690 | 0.4511 | 0.2143 | 0.4332 |
| | ✓ | 0.5722 | 0.5984 | 0.6305 | 0.5796 | 0.5869 | 0.8086 | 0.6558 | 0.6080 | 0.6112 | 0.5920 | 0.6158 | 0.5653 | 0.6187 |
| VAD-Qwen3-4B | × | 0.3350 | 0.4475 | 0.5231 | 0.3988 | 0.3745 | 0.8513 | 0.4768 | 0.4947 | 0.5704 | 0.4915 | 0.4195 | 0.2910 | 0.4728 |
| | ✓ | 0.9183 | 0.7278 | 0.7044 | 0.6071 | 0.6085 | 0.7999 | 0.5516 | 0.5516 | 0.6778 | 0.6333 | 0.5663 | 0.5768 | 0.6574 |
| VAD-Qwen3-8B | × | 0.3433 | 0.4176 | 0.4933 | 0.3378 | 0.4195 | 0.7760 | 0.5076 | 0.5296 | 0.5168 | 0.5140 | 0.4328 | 0.3061 | 0.4662 |
| | ✓ | 0.9219 | 0.7329 | 0.7104 | 0.6140 | 0.5952 | 0.8090 | 0.6727 | 0.5387 | 0.7012 | 0.6480 | 0.5819 | 0.5938 | 0.6767 |
| VAD-Qwen3-30B | × | 0.3561 | 0.4387 | 0.5066 | 0.3580 | 0.4254 | 0.7880 | 0.5374 | 0.5493 | 0.5567 | 0.6939 | 0.5626 | 0.3291 | 0.5086 |
| | ✓ | **0.9389** | **0.7384** | **0.7246** | **0.6282** | **0.6305** | **0.8191** | **0.6837** | **0.6594** | **0.7921** | **0.7287** | **0.7140** | **0.6553** | **0.7261** |

**Table 5**
Modality ablation study with fixed Qwen3-0.6B using LoRA fine-tuning setup on Task 2 (30 s clip level). V represents video, A represents audio, and D represents dialogue transcripts.

| Setting | A2 | A5 | A6 | A7 | A8 | B1 | B3 | B5 | B9 | B10 | B11 | B12 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V only | 0.4460 | 0.3354 | 0.5700 | 0.5260 | 0.3799 | 0.5594 | 0.5158 | 0.5135 | 0.6186 | 0.5741 | 0.6213 | 0.3919 | 0.5044 |
| A only | 0.4653 | 0.4079 | 0.4690 | 0.4410 | 0.4066 | 0.3519 | 0.3501 | 0.6815 | 0.3373 | 0.3914 | 0.7315 | 0.2973 | 0.4443 |
| D only | 0.5034 | 0.3575 | 0.5608 | 0.5318 | 0.4290 | 0.3272 | 0.3795 | 0.7053 | 0.3584 | 0.4690 | 0.7310 | 0.3037 | 0.4714 |
| V + A | 0.3882 | 0.3386 | 0.6131 | **0.6952** | **0.6255** | 0.6351 | 0.6167 | 0.6700 | 0.5828 | **0.7310** | 0.6434 | 0.4313 | 0.5805 |
| V + D | 0.3946 | 0.3217 | **0.7127** | 0.5796 | 0.5474 | 0.6163 | 0.5309 | 0.6126 | **0.6714** | 0.5929 | 0.6291 | 0.3827 | 0.5493 |
| A + D | 0.3005 | 0.3836 | 0.5626 | 0.4520 | 0.4910 | 0.4745 | 0.4194 | **0.7356** | 0.4107 | 0.4768 | **0.7838** | 0.3478 | 0.4866 |
| V + A + D | **0.5722** | **0.5984** | 0.6305 | 0.5796 | 0.5869 | **0.8086** | **0.6558** | 0.6080 | 0.6112 | 0.5920 | 0.6158 | **0.5653** | **0.6187** |

**Table 6**
Modality ablation study with fixed Qwen3-0.6B using LoRA fine-tuning setup on Task 1 (10-category paradigm classification). V represents video, A represents audio, and D represents dialogue transcripts.

| Setting | V | A | D | Acc |
|---|---|---|---|---|
| V only | ✓ | | | 0.4636 |
| A only | | ✓ | | 0.3196 |
| D only | | | ✓ | 0.5631 |
| V + A | ✓ | ✓ | | 0.4408 |
| V + D | ✓ | | ✓ | 0.5573 |
| A + D | | ✓ | ✓ | 0.5243 |
| V + A + D | ✓ | ✓ | ✓ | **0.6119** |

63.05% (VAD-Qwen3-30B with LoRA). These results suggest that multimodal alignment effectively captures fine-grained social signals that are difficult to model using unimodal representations. Similarly, for the B-series items related to communicative behaviors and repetitive patterns, consistent improvements are observed in 6 out of 7 items (except for B1). Large gains are achieved on B3 (+38.69%) and B12 (+33.02%), indicating that integrating visual-audio-dialogue cues with LLMs reasoning enhances structured behavioral assessment. Overall, the improvements are not confined to a specific subset of items but span different behavioral dimensions, demonstrating the robustness and generalizability of the proposed multimodal framework.

Similar to Task 1, model scaling consistently improves item-level prediction accuracy. These results verify that combining aligned visual-audio-dialogue features with LLMs significantly enhances detailed item-level assessment.

### 5.4. Ablation studies

#### 5.4.1. Analysis of session-level label propagation

To investigate whether session-level item labels provide consistent supervision across different interaction contexts, we further analyzed item-level performance across ADOS paradigms. As shown in Table 7, some items are strongly associated with their clinically designed paradigms, while others can be predicted across multiple interaction contexts. For example, B10 and B11 perform well in the paradigms designed to elicit joint attention-related behaviors, while B1 and B9 remain predictable across several paradigms. These findings suggest that although session-level label propagation introduces some clip-level uncertainty, multiple interaction contexts can provide complementary behavioral evidence.

To investigate the contribution of each modality, we conduct modality ablation experiments on both Task 1 and Task 2. In this study, the backend LLM is fixed as Qwen3-0.6B with LoRA fine-tuning, and we vary the input modality combinations.

#### 5.4.2. Modality contribution analysis

**Task 1: Paradigm Classification.** As shown in Table 6, the text-only setting achieves the best performance among unimodal configurations (56.31%), outperforming visual-only (46.36%) and audio-only (31.96%). This indicates that dialogue transcripts carry strong semantic cues for paradigm identification.

Interestingly, combining modalities does not always guarantee improvement. For example, the V+A setting (44.08%) performs worse than visual-only (46.36%). This counterintuitive result indicates that audio features alone provide limited discriminative information compared with dialogue transcripts for paradigm classification. Since Task 1 primarily requires understanding interaction structures, low-level acoustic cues such as intonation and pitch variations may introduce less information when not grounded in dialogue transcripts. When fused with visual features without explicit modal weighting, audio cues may dilute the dominant visual cues, leading to a slight performance decline. In contrast, combining visual and dialogue transcript modalities (V+D) yields significant improvements, indicating that dialogue-level text information plays a crucial role in paradigm recognition.

Overall, the full-modality configuration (V+A+D) demonstrated the highest accuracy rate of 61.19%, proving that complementary information can be generated when three modalities collaborate in modeling.

**Task 2: Item-level Behavioral Assessment.** The modality ablation results for Task 2 are presented in Table 5. Among single-modality configurations, the visual modality achieves the best performance (50.44%). The V+A combination yields superior results (58.05%) compared to other bimodal settings, while the full multimodal configuration (V+A+D) achieves the highest overall performance (61.87%).

Notably, the V+A combination (58.05%) outperforms both V+D (54.93%) and A+D (48.66%). This observation suggests that

**Table 7**
Item-level accuracy (%) of VAD-Qwen3-30B on test clips from each ADOS paradigm.

| Paradigm | A2 | A5 | A6 | A7 | A8 | B1 | B3 | B5 | B9 | B10 | B11 | B12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FR | 94.23 | 74.65 | 78.17 | 65.12 | 68.13 | 84.07 | 66.62 | 65.37 | 79.17 | 76.29 | 73.65 | 64.12 |
| RN | 86.67 | 73.33 | 73.33 | 86.67 | 80.00 | 93.33 | 80.00 | 93.33 | 93.33 | 66.67 | 53.33 | 46.67 |
| RJA | 91.89 | 83.78 | 56.76 | 72.97 | 32.43 | 78.38 | 70.27 | 56.76 | 81.08 | 62.16 | 76.76 | 75.68 |
| BUP | 92.64 | 72.39 | 64.42 | 52.15 | 55.21 | 79.75 | 77.30 | 67.48 | 83.44 | 72.39 | 66.26 | 70.55 |
| ARO | 95.56 | 76.67 | 68.33 | 58.89 | 60.00 | 86.11 | 70.00 | 61.67 | 82.22 | 68.33 | 71.67 | 68.89 |
| ASR | 92.31 | 75.64 | 67.95 | 69.23 | 43.59 | 79.49 | 69.23 | 58.97 | 82.05 | 65.38 | 76.92 | 66.67 |
| FSI | 96.58 | 68.49 | 67.81 | 69.18 | 63.01 | 81.51 | 76.03 | 67.81 | 78.08 | 70.55 | 73.29 | 67.81 |
| BIP | 93.84 | 76.11 | 74.38 | 63.79 | 63.79 | 82.02 | 69.95 | 67.00 | 80.79 | 74.38 | 68.47 | 67.00 |
| SN | 92.16 | 69.47 | 67.79 | 57.42 | 62.75 | 76.19 | 61.62 | 68.07 | 73.11 | 69.47 | 72.27 | 61.62 |

fine-grained behavioral assessment relies more heavily on audio-visual interaction cues rather than purely textual semantics. The ADOS assessment items encompass vocal prosody (A2, A5), nonverbal engagement (A6, A7, A8), eye contact (B1), facial expressions (B3), social interaction (B5, B9, B12), and attention-related behaviors (B10, B11). These characteristics are more effectively captured through visual and audio signals. In contrast, the A + D configuration demonstrates relatively weaker performance. Since audio features and dialogue transcripts are highly correlated, their combination may introduce redundancy without providing sufficiently complementary behavioral information. Moreover, the absence of visual cues limits the model's ability to capture nonverbal interaction patterns, which are essential for accurate scoring of many ADOS items.

At the item level, substantial variability in modality preference is observed. The V + A configuration achieves strong performance on A7 (Pointing), A8 (Gestures), and B10 (Spontaneous Initiation of Joint Attention), indicating that audiovisual interaction cues are critical for evaluating nonverbal engagement behaviors. In contrast, the V + D combination performs particularly well on interaction-related items such as A6 (Using Another's Body to Communicate) and B9 (Showing), highlighting the essential role of conversational semantics in socially interactive contexts. Meanwhile, the superior performance of the A + D configuration on B5 (Shared Enjoyment in Interaction) and B11 (Response to Joint Attention) can be attributed to the intrinsic reliance of these items on dialogue transcripts and acoustic cues. Both items primarily assess whether the child expresses emotional involvement and provides appropriate responses within conversational settings. Such characteristics are more effectively captured through dialogue transcripts and acoustic cues than through visual motion patterns alone.

Overall, different behavioral dimensions exhibit distinct modality dependencies, and the complete multimodal configuration provides the most robust and generalized performance across diverse ADOS items.

#### 5.4.3. Analysis of modality-weighted fusion

To further investigate whether explicit modality weighting can improve the original concatenation-based fusion strategy, we compared the standard concatenation approach with two lightweight gating mechanisms: global scalar gating and sample-adaptive gating. All experiments used the same Qwen3-0.6B LLM, visual-audio-dialogue inputs, subject-independent data split, and training configuration as in the main experiments.

The global scalar gate learns a fixed set of weights for the three modalities, while the sample-adaptive gate dynamically generates modality weights based on the multimodal representation of each input sample. Both gating mechanisms are applied after modality-specific representations are extracted and before the features are fed into the LLM.

As shown in Table 8, both gating strategies improve upon the original concatenation-based fusion for paradigm classification. The results indicate that explicit modality weighting can reduce potential modality competition and provide more effective multimodal representations for paradigm classification.

For ADOS item-level prediction, the improvement is relatively limited, although the sample-adaptive gate achieves the best performance (63.62%). This suggests that modality-level weighting can provide additional information for fine-grained behavioral assessment, but it cannot fully model the complex relationships among different modalities. More advanced cross-modal modeling methods remain an important direction for future work.

**Table 8**
Comparison of the original concatenation fusion and lightweight modality-gating alternatives. Task 1 is 10-category paradigm classification; Task 2 is mean accuracy across the 12 ADOS items.

| Fusion method | Task 1 accuracy (%) | Task 2 mean accuracy (%) |
|---|---|---|
| Concat | 61.19 | 61.87 |
| Global scalar gate | **66.55** | 62.52 |
| Sample-adaptive gate | 66.23 | **63.62** |

**Table 9**
Visual–audio fusion analysis for Task 1. Visual and Concat results are reused from the original ablation study, while gating experiments use Qwen3-0.6B under the same split and metric.

| Input | Fusion | Accuracy (%) |
|---|---|---|
| Visual | – | 46.36 |
| Visual + audio | Concat | 44.08 |
| Visual + audio | Global scalar gate | 55.61 |
| Visual + audio | Sample-adaptive gate | 57.40 |

**Table 10**
Controlled component ablation with a fixed LLM (Qwen3-0.6B), visual–audio–dialogue inputs, subject-independent data split, and evaluation metric. Complete instruction denotes the original clinically oriented instruction format; minimal instruction denotes a single direct task prompt without the clinical role description or additional task prompts. Frozen random adapters are randomly initialized visual and audio projections that are not updated during training.

| Setting | Instruction design | LLM update | Visual/audio adapters | Task 2 mean accuracy (%) |
|---|---|---|---|---|
| E1 | Minimal | Frozen | Trainable | 40.06 |
| E2 | Complete | Frozen | Trainable | 43.32 |
| E3 | Complete | LoRA | Trainable | **61.87** |
| E4 | Complete | LoRA | Frozen random | 41.53 |

To further investigate the performance drop observed in the visual–audio concatenation setting, we compared visual-only input and direct visual–audio concatenation with global scalar gating and sample-adaptive gating. As shown in Table 9, directly concatenating visual and audio features reduced Task 1 accuracy from 46.36% to 44.08%. In contrast, the global scalar gate and sample-adaptive gate achieved accuracies of 55.61% and 57.40%, respectively. These results suggest that the performance decrease with direct concatenation does not necessarily mean that audio information is unhelpful, but may be caused by the lack of effective modality balancing. Learned modality weighting allows the model to better utilize information from both visual and audio modalities.

#### 5.4.4. Component-wise ablation of instruction, LoRA, and adapter

To understand the contributions of different optimization components, we further conducted controlled ablation experiments by separately evaluating the effects of clinical instruction design, LoRA-based LLM adaptation, and learnable multimodal adapters.

As shown in Table 10, replacing the minimal instruction with the clinically oriented instruction improves the mean item prediction accuracy from 40.06% to 43.32%, a gain of 3.26 percentage points, indicating that a more complete clinical task description provides clearer task constraints. Introducing LoRA-based LLM adaptation further increases the accuracy to 61.87%, an improvement of 18.55 percentage points over the frozen-LLM setting and the largest gain among the evaluated components. When the learnable visual and audio adapters are replaced with frozen random projections, the accuracy decreases to 41.53%, which is 20.34 percentage points lower than the complete model. These results show that LoRA improves adaptation of the pretrained LLM to ADOS item prediction, while learnable adapters are important for aligning visual and audio features with the LLM representation space.

### 5.5. Subject-level ASD detection performance

After obtaining ADOS item-level scores for all video clips, we further evaluated subject-level ASD detection performance. The training set is the same as in the item-level assessment, which includes 402 subjects

**Table 11**
Subject-level ASD detection performance under physician annotations.

| Model | Rule | LR | SVM | NB |
|---|---|---|---|---|
| VIT [32] | 0.6222 | 0.7333 | 0.7333 | 0.5333 |
| Whisper-Large-V3 [53] | 0.5777 | 0.7333 | 0.7333 | 0.6222 |
| BERT [13] | 0.6000 | 0.7333 | 0.7333 | 0.5333 |
| MULT [60] | 0.6444 | 0.6667 | 0.7111 | 0.5778 |
| SCBU-Qwen3-Omni [37] | 0.7333 | 0.7333 | 0.7333 | 0.6200 |
| VAD-Qwen2.5-0.5B | 0.6444 | 0.7333 | 0.7111 | 0.6667 |
| VAD-Qwen3-0.6B | 0.6666 | 0.7556 | 0.7333 | 0.6000 |
| VAD-Qwen3-4B | 0.6444 | 0.7333 | 0.7556 | 0.6000 |
| VAD-Qwen3-8B | 0.7555 | 0.7556 | 0.7778 | 0.6889 |
| VAD-Qwen3-30B | 0.7555 | 0.7556 | **0.8000** | 0.6667 |

**Table 12**
Subject-level ASD detection performance under ADOS score rule.

| Model | Rule | LR | SVM | NB |
|---|---|---|---|---|
| VIT [32] | 0.6000 | 0.6444 | 0.6889 | 0.3556 |
| Whisper-Large-V3 [53] | 0.4222 | 0.6000 | 0.6000 | 0.3778 |
| BERT [13] | 0.5555 | 0.7333 | 0.7333 | 0.3556 |
| MULT [60] | 0.4222 | 0.5778 | 0.6000 | 0.4000 |
| SCBU-Qwen3-Omni [37] | 0.7333 | 0.7333 | 0.7333 | 0.7333 |
| VAD-Qwen2.5-0.5B | 0.7111 | 0.7778 | 0.7333 | 0.6889 |
| VAD-Qwen3-0.6B | 0.7111 | 0.8000 | 0.7333 | 0.6889 |
| VAD-Qwen3-4B | 0.8000 | 0.8222 | 0.8222 | 0.7556 |
| VAD-Qwen3-8B | 0.8444 | 0.8000 | 0.8444 | 0.8000 |
| VAD-Qwen3-30B | **0.8667** | 0.8000 | **0.8667** | 0.8222 |

(283 ASD and 119 TD; 328 males and 74 females), while the test set includes 45 subjects (33 ASD and 12 TD; 38 males and 7 females).

#### 5.5.1. Ground truth settings

We evaluated ASD detection under two different ground-truth (GT) settings.

(1) **Physician Annotations.** This GT was determined by experienced doctors. The final diagnosis was reached through consensus between at least two senior clinical experts, based on a comprehensive assessment of ADOS-2 scores, clinical observations, and developmental history. The physician-annotated GT contains two categories: *ASD* and *TD*.
(2) **ADOS Rule-Based Annotations.** This GT was generated according to the ADOS-2 Module 1 scoring algorithm. This GT contains three categories: *Autism*, *Autism Spectrum*, and *Non-Spectrum*. For evaluation under physician annotations, the categories *Autism* and *Autism Spectrum* were merged into a single *ASD* class.

#### 5.5.2. ASD detection strategies

Based on the predicted ADOS item-level assessment scores, we implemented two ASD detection strategies:

(1) **ADOS Rule-Based Inference (Rule).** The final diagnostic result was determined directly using the ADOS assessment algorithm [20].
(2) **Machine Learning-Based Classification.** We further trained three machine learning classifiers on the training set to learn diagnostic decision boundaries from predicted ADOS scores:
   - Logistic Regression (LR)
   - Support Vector Machine (SVM)
   - Naive Bayes (NB)

   The trained classifiers were then applied to the test set to obtain subject-level predictions.

#### 5.5.3. Results under physician annotations

The subject-level ASD detection results under physician annotations are presented in Table 11. Our multimodal approaches, implemented with different LLMs, consistently outperform feature-based deep learning methods and script-centric behavior understanding methods. Among all methods, VAD-Qwen3-30B combined with a support vector machine (SVM) achieves the highest accuracy (80.00%), followed by VAD-Qwen3-8B (77.78%). This indicates that large instruction-tuned multimodal models are capable of generating more discriminative ADOS scores for ASD classification. This suggests that learned decision boundaries are more robust to prediction noise at the item level compared to deterministic ADOS rules. The improvement is especially pronounced for methods with relatively lower ADOS scoring accuracy (e.g., VIT + SVM improves by 11.11% over VIT + Rule).

#### 5.5.4. Results under ADOS rule-based labels

Table 12 reports the validation results based on ground truth derived from ADOS assessment algorithm. VAD-Qwen3-30B achieves the best overall performance, reaching 86.67% under both rule-based and SVM settings. Compared to the results under physician annotations, accuracy is generally higher when evaluated against ADOS rule-derived ground truth considering it is a 3-category classification. This outcome is expected, as rule-based ground truth strictly follows ADOS cutoff thresholds. Therefore, accurate ADOS item prediction can effectively support subject-level ASD detection under this evaluation protocol. In contrast, physician diagnosis incorporates broader clinical judgment beyond standardized scoring, and relying solely on ADOS score prediction is insufficient to fully capture diagnostic complexity.

Overall, these findings demonstrate that precise ADOS item prediction effectively supports subject-level ASD detection, and the integration of machine learning classifiers further enhances diagnostic robustness.

#### 5.5.5. Robustness analysis with five-fold cross-validation

To further evaluate the robustness of subject-level ASD detection, we performed subject-independent stratified five-fold cross-validation on all 447 participants. All samples from the same participant were assigned to the same fold to prevent subject-level leakage, and the diagnostic distribution was preserved as closely as possible across folds. In each fold, the SVM diagnostic backend was fitted only on the training subjects and evaluated on the held-out subjects. We selected VAD-Qwen3-0.6B and compared it with other baselines under the same cross-validation setting.

As shown in Fig. 7, VAD-Qwen3-0.6B achieved the highest OOF accuracy of 68.18% with a 95% bootstrap confidence interval of 63.70–71.87%. This result exceeds those of ViT (64.36%), MULT (64.14%), BERT (63.09%), and Whisper-Large-V3 (62.87%). VAD-Qwen3-0.6B also achieved the lowest standard deviation across folds at 2.58 percentage points, compared with 6.64 for ViT, 4.54 for MULT, 3.20 for BERT, and 3.66 for Whisper-Large-V3, indicating more consistent performance across subject splits.

The five-fold cross-validation accuracy is more conservative than the 73.33% obtained under the original fixed split. This analysis does not replace the original evaluation, but provides complementary evidence that the effectiveness of the proposed framework is not limited to a particular subject split.

#### 5.5.6. Confusion matrix analysis

Fig. 6 illustrates the confusion matrices of the best-performing model (VAD-Qwen3-30B with SVM) under the two ground-truth settings.

Under physician annotations (binary classification), most misclassifications occur when ASD cases are incorrectly predicted as TD, accounting for eight instances, whereas only one TD case is misclassified as ASD. This asymmetric error pattern suggests that the model tends to adopt a conservative strategy when predicting ASD. Such behavior may arise because some ASD cases exhibit relatively mild behavioral symptoms, making them difficult to distinguish from TD individuals based solely on clip-level ADOS item predictions. Moreover, since both Autism and Autism Spectrum cases are merged as ASD in the physician annotations, the system might predict Autism Spectrum as TD. Under the ADOS rule-based annotations (three-class classification), the

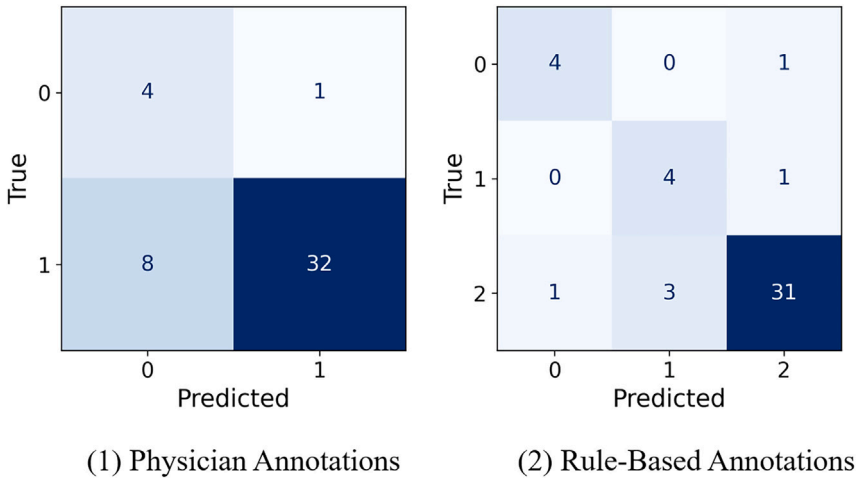



**Fig. 6.** The confusion matrices of the best-performing model (VAD-Qwen3-30B with SVM) for subject-level ASD detection under the two ground-truth settings.

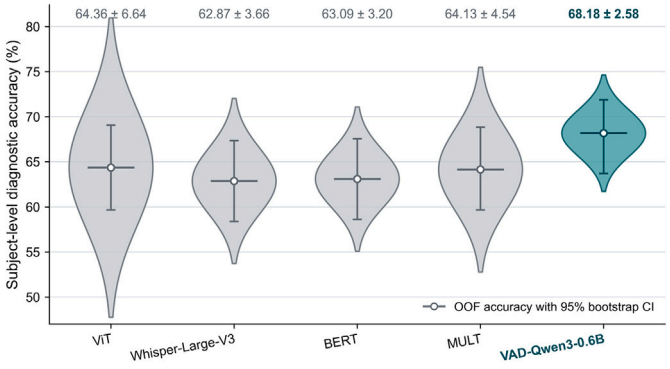



**Fig. 7.** Five-fold stratified subject-independent cross-validation for subject-level ASD detection using the SVM backend. The shaded profiles visualize the reported fold mean and standard deviation, while the circles and vertical error bars indicate the OOF accuracy and its 95% bootstrap confidence interval, respectively. OOF denotes pooled out-of-fold predictions.

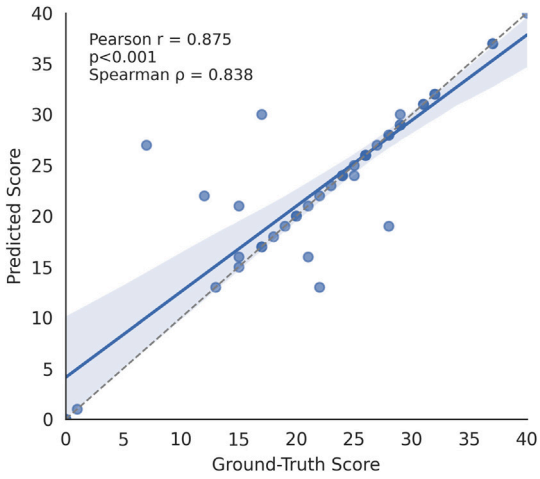



**Fig. 8.** Correlation analysis between predicted and doctor's item-level scores. Each point represents one subject, where the predicted score is obtained by aggregating the 12 item-level predictions. The solid line represents the fitted regression, with the shaded region indicating the confidence interval, while the dashed diagonal line ($y = x$) denotes perfect agreement.

confusion matrix shows that most predictions lie along the diagonal, indicating strong consistency between the predicted scores and the rule-based diagnostic categories. The primary confusion occurs between the categories of Autism and Autism Spectrum. This pattern is consistent with clinical observations, as the boundary between Autism and Autism Spectrum can be ambiguous and often depends on the severity of social and communication impairments.

Overall, the confusion matrices indicate that the proposed ASD detection strategy successfully captures the major diagnostic patterns reflected in ADOS assessments while maintaining strong agreement with both doctor judgments and rule-based diagnostic criteria.

*5.5.7. Correlation analysis*

To evaluate the consistency between the predicted item-level assessment scores and the Doctor's item-level assessment scores, we perform a correlation analysis. As shown in Fig. 8 each point represents one subject in the test set, where the predicted score is obtained by aggregating the 12 item-level assessment scores. The results demonstrate a strong positive correlation between the predicted scores and the ground-truth scores, with a Pearson correlation coefficient of $r = 0.875$ ($p < 0.001$) and a Spearman correlation coefficient of $\rho = 0.838$. The high Pearson coefficient indicates a strong linear relationship, while the Spearman coefficient further confirms that the monotonic trends are well preserved.

As shown in Fig. 8, most samples are distributed around the diagonal reference line ($y = x$), indicating that the predicted scores are well aligned with the ground-truth ratings across different severity levels. Moreover, predictions in the high-score range exhibit tighter clustering near the diagonal, suggesting that the model achieves more stable performance for subjects with more pronounced ASD-related behavioral characteristics.

Several outliers can be observed where the predicted scores deviate substantially from the doctor ratings. For instance, one sample in the upper-left region shows a predicted total score of 27 compared to a doctor rating of 7. After reviewing the video, we found that the child frequently moved out of the camera view during recording, leading to incomplete capture of observable behaviors and consequently an overestimation of the predicted score. This observation highlights the impact of missing or degraded modality information and suggests that handling modality absence or using multi-view cameras should be considered in future work.

Overall, these results demonstrate that the proposed framework is capable of capturing ADOS-related behaviors, providing strong evidence for its clinical consistency and interpretability.

**Table 13**
Subject-level ASD detection accuracy under three clip filtering settings (No Filtering, Automatic Filtering, and Manual Filtering).

| Method | Accuracy (%) |
|---|---|
| No filtering | 77.78 |
| Automatic filtering | 82.22 |
| Manual filtering | 86.67 |

### 5.6. Fully automated assessment of long videos

After separately validating the model performance for Task 1 and Task 2, we proceed to evaluate fully automated assessment on long videos. To this end, we integrate Task 1 and Task 2 into a unified pipeline: the Task 1 model is first used to filter out non-paradigm clips from long videos, followed by the Task 2 model to generate item-level scores. Finally, the ADOS rule-based ASD detection strategy is applied to produce the final diagnostic outcome.

To evaluate the performance of fully automated long-video assessment, we design the following experimental settings: (1) No Filtering: ADOS multi-item behavioral assessment and ASD detection are performed directly on the full videos without excluding non-paradigm clips. (2) Automatic Filtering (Task 1 + Task 2 pipeline): Non-paradigm clips are excluded based on the paradigm categories predicted by the Task 1 model, followed by ADOS multi-item behavioral assessment and ASD detection. This setting evaluates the effectiveness of combining Task 1 and Task 2. (3) Manual Filtering (Task 2): Non-paradigm clips are excluded using manually annotated timestamps, after which ADOS multi-item behavioral assessment and ASD detection are performed. This setting follows the same procedure as Task 2 and serves as an upper bound for performance.

To ensure fair comparison, all experiments use the best-performing model from Task 2, VAD-qwen3-30B. The Task 1 model used in Setting (2) is also its best-performing model, VAD-qwen3-30B.

As shown in Table 13, filtering out non-paradigm segments significantly improves ASD detection performance. Compared to the no-filtering setting (77.78%), automatic filtering based on the Task 1 model increases the accuracy to 82.22%, demonstrating both the effectiveness and necessity of the Task 1 design. However, due to errors in paradigm classification, the Task 1 model still limits the overall detection performance, resulting in a performance gap between automatic and manual filtering. In future work, integrating Task 1 and Task 2 into a unified VAD framework for joint training may simplify the assessment process while further enhancing overall performance.

### 5.7. Discussions and limitations

Our method evaluates ADOS-M1 behavioral scores using a multimodal large language model framework. In practice, ADOS clinical recordings are typically lengthy and contain multiple interaction paradigms. To address this, we first define a paradigm classification task to identify the relevant paradigm clips requiring assessment, thereby filtering out irrelevant video clips. Next, we formulate an item-level behavioral assessment task based on the ADOS scoring rules, where LLMs analyze multimodal cues to generate ADOS-M1 item-level results that can support clinical diagnosis. Finally, the predicted behavioral scores are used to perform end-to-end ASD detection through both rule-based ADOS inference and machine learning–based classifiers.

Although the proposed method achieves strong performance across tasks, several limitations remain for real-world deployment. First, clinical diagnostic videos must be recorded in relatively controlled environments with synchronized audio and video to ensure sufficient data quality for behavioral assessment. Second, although lightweight modality adapters and LoRA reduce the demand for large training datasets, autism research [2,39,73] often involves data from limited sample sizes (typically fewer than 100 subjects), which may significantly affect the robustness of the proposed framework. Finally, the current framework does not explicitly model the temporal interactions between video and audio signals, which may limit the ability of LLMs to capture the consistency between behavioral actions and speech cues.

Recent advances in multimodal learning, including spatiotemporal representation learning [7,66–68] and multimodal adaptation strategies [40,55,61], provide promising directions for improving ADOS behavioral assessment. Incorporating these techniques may help capture complex behavioral dependencies and enhance multimodal alignment, which we leave for future investigation.

Our cohort was collected from a single urban medical center and consists of Mandarin-/Cantonese-speaking children, which may limit cross-cultural generalizability. Socioeconomic variables were not retained due to privacy regulations. In addition, the dataset exhibits class and demographic imbalance (e.g., ASD vs. TD, male vs. female). Future work will evaluate the trained model on independent cohorts from multiple clinical centers to assess its generalizability across broader regional, linguistic, and recording conditions. Although only Viewpoint 1 was used in this study, quality-aware view selection and multi-view fusion will be explored to address occlusion and out-of-view behavior. Future data collection will also retain independent item-level ratings from multiple ADOS-qualified clinicians before consensus, enabling inter-rater reliability analysis using weighted Cohen's kappa and confidence intervals.

## 6. Conclusion

This paper proposes a multimodal large language model framework for automated ADOS-M1 behavioral assessment. The framework integrates visual, acoustic, and dialogue transcript information through modality-specific adapters and leverages instruction fine-tuning with LoRA to improve the understanding of multimodal cues in ADOS assessment. By reformulating ADOS paradigm classification and item-level behavioral assessment as structured text generation tasks, the framework enables LLMs to reason over multimodal behavioral cues in accordance with ADOS scoring protocols. Experimental results show that the proposed framework outperforms unimodal models, conventional multimodal fusion methods, and script-centric behavioral understanding approaches. Moreover, the predicted ADOS scores enable reliable subject-level ASD detection, achieving up to 86.67% accuracy under doctor-annotated ground truth. These results demonstrate that multimodal LLMs provide an interpretable and clinically compatible approach for ADOS behavioral assessment. Future research will focus on enhancing the perception capabilities for multimodal temporal information and exploring the generalization capabilities of datasets across ADOS modules.

## CRediT authorship contribution statement

**Wenxing Liu:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Yueran Pan:** Investigation, Data curation. **Dong Zhang:** Validation, Conceptualization. **Hongzhu Deng:** Validation, Project administration, Data curation. **Xiaobing Zou:** Validation, Resources, Funding acquisition, Conceptualization. **Ming Li:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. ADOS items

Table 14 lists the ADOS-M1 items evaluated in this study, including the item abbreviations, full item names, and their corresponding brief descriptions.

**Table 14**
The item abbreviations (Item), full item names (Name), and corresponding brief descriptions (Description) of the ADOS items.

| Item | Name | Description |
|---|---|---|
| A2 | Frequency of Vocalizations Directed to Others | This item evaluates the number of vocalizations that are socially directed toward others. |
| A5 | Stereotyped or Idiosyncratic Use of Words or Phrases | This item includes delayed echolalia or other highly repetitive utterances with consistent intonation patterns. |
| A6 | Use of Another Person's Body to Communicate | This item focuses on situations in which a child uses another person's body as a tool to accomplish an action. |
| A7 | Pointing | This item describes socially directed pointing gestures used to request or share attention. |
| A8 | Gestures | This item evaluates the use of emotional or descriptive gestures other than pointing. |
| B1 | Unusual Eye Contact | This item distinguishes appropriate, flexible, and socially regulated eye gaze from gaze that is limited in flexibility, appropriateness, or contextual use. |
| B3 | Facial Expressions Directed to Others | This item evaluates whether the child's facial expressions are directed toward another person to communicate emotion. |
| B5 | Shared Enjoyment in Interaction | This item assesses the child's social responses during activities or interaction-eliciting stimuli. |
| B9 | Showing | Showing is defined as intentionally orienting or placing an object so that another person can see it, without a clear purpose of requesting help or participating in a routine. |
| B10 | Spontaneous Initiation of Joint Attention | This item evaluates attempts by the child to draw another person's attention to an object that neither individual is touching. |
| B11 | Response to Joint Attention | This item evaluates the child's response when the examiner directs the child's attention to a distant object using gaze and/or pointing. |
| B12 | Quality of Social Overtures | This summary item focuses on the quality of the child's attempts to initiate social interaction rather than their frequency. |

## Appendix B. Instruction data

### *B.1. Task 1*

Instructional data can enhance an LLM's capability to handle downstream tasks. Therefore, we design the instructional data for paradigm classification comprising system prompt and question, as detailed in Fig. 9. The system prompts consist of four parts: (1) Input Specifications define the input format. (2) Task clarifies the task content. (3) Paradigm Introduction details the workflow and characteristics of each paradigm. (4) Output Specifications constrain the format of the output results.

## System prompt

### Input Specifications ###
You will receive an embedding vector for a 30-second video clip, the corresponding embedding vector for its audio track, and the associated conversation content.

### Task ###
Your task is to accurately identify which AODS-M1 paradigm this clip belongs to.

### Paradigm Introdution ###
- Free: Doctors provide a range of age-appropriate toys for children to play with in a natural and relatively unstructured manner. Specifically, during the paradigm process, a large number of toys are placed on the table or floor.
- Response to name: Doctors or parents call out the child's name during the activity. Specifically, the parent/doctor walks behind the child with a microphone and calls out the child's name.
- Response to joint attention: Doctors attempt to guide children to engage in joint attention through eye contact, gestures, or verbal cues. Specifically, a rabbit toy appears during the paradigm.
- Bubble play: The doctor encourages the child to participate in the interaction by blowing bubbles. Specifically, bubble solution is used during the paradigm process.
- Anticipation of a routine with objects: Doctors assess children's ability to predict the next action in a familiar sequence of object manipulations. Specifically, balloons appear during the paradigm.
- Anticipation of a social routine: Observe children for signs of anticipating the next action in familiar social interactions. Specifically, doctors play games with children, such as peek-a-boo, tickling, and lifting them up high.
- Functional and symbolic imitation: The doctor demonstrates familiar movements from everyday life. Specifically, frogs, cars, cups, cylinders, or flowers appear during the modeling process.
- Birthday party: The doctor simulates a birthday party scenario. Specifically, props needed for a birthday party, such as dolls, plates, knives, forks, candles, cups, and towels, appear during the paradigm process.
- Snack: The doctor provides the child with food or snacks and observes the child's behavior in asking others for food or reacting to food. Specifically, two snack boxes appear during the paradigm process.
- no paradigm: Interactive behaviors other than the above nine paradigms
- invalid video: Doctors did not interact effectively with children. Specifically, there may have been a shortage of staff and chaos.

### Output Specifications ###
Your output should be a AODS-M1 patterns present in the audio-visual clip. The paradigm category only includes the following: Free, Response to name, Response to joint attention, Bubble play, Anticipation of a routine with objects, Anticipation of a social routine, Functional and symbolic imitation, Birthday party, Snack, no paradigm, invalid video.

## Question

What are the ADOS-M1 paradigm type in this audio-visual clip?

**Fig. 9.** The instruction data of paradigm classification.

### B.2. Task 2

We design the instruction data for the ADOS item score prediction, including system prompts and questions, as detailed in Fig. 10. The system prompts comprise three sections: (1) Input Specifications define the input format. (2) Task clarifies the task content. (3) Output Specifications constrain the output format. The questions comprise 12 distinct ADOS items. For each item's scoring assessment, we designed it in a multiple-choice format, querying each item sequentially through a multi-round dialogue process.

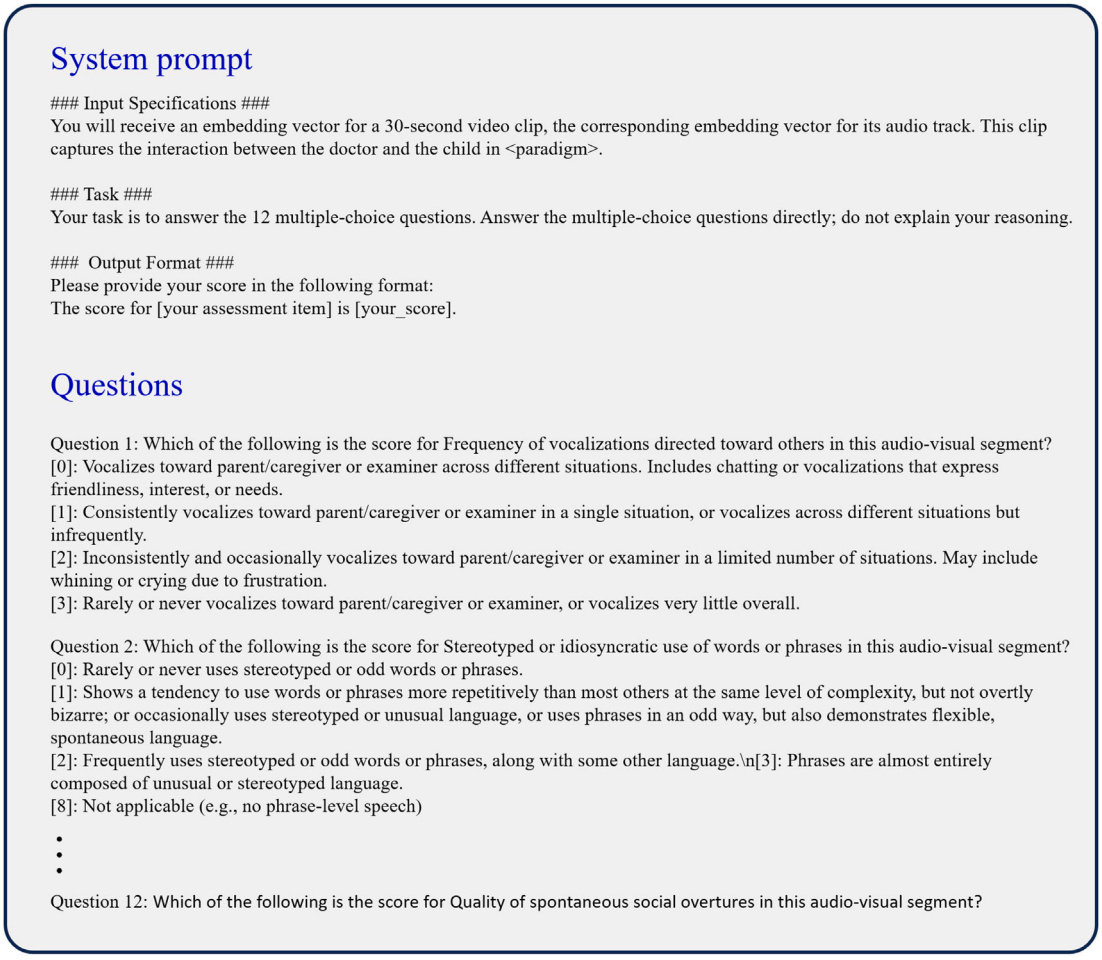



**Fig. 10.** The instructions for the ADOS item score prediction.

## Data availability

The authors do not have permission to share data.

## References

[1] R. AlSaad, et al., Multimodal large language models in health care: applications, challenges, and future outlook, J. Med. Internet Res. 26 (2024) e59505.
[2] J.R. Asmetha, R. Senthilkumar, DSVTN-ASD: detection of stereotypical behaviors in individuals with autism spectrum disorder using a dual self-supervised video transformer network, Neurocomputing 624 (2025) 129397.
[3] R. Assaf, Z. Shehabeddine, V. Ramesh, Screening autism spectrum disorder in children using machine learning on speech transcripts, Sci. Rep. 15 (2025) 34134.
[4] M. Bain, J. Huh, T. Han, A. Zisserman, Whisperx: time-accurate speech transcription of long-form audio, arXiv preprint arXiv:2303.00747, 2023.
[5] F. Briend, C. David, S. Silleresi, J. Malvy, S. Ferré, M. Latinus, Voice acoustics allow classifying autism spectrum disorder with high accuracy, Transl. Psychiatr. 13 (2023) 250.
[6] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, B. Schuller, Speech-based age and gender prediction with transformers, in: Speech Commun, 2023, pp. 46–50.
[7] H. Chen, L. Yan, Q. Wang, J. Zhang, F. Guan, P. Li, Chain-of-search: parameter-efficient reasoning for zero-shot object navigation, in: Proc. AAAI Conf. Artif. Intell. (AAAI), 2026, pp. 1721–1729.
[8] F. Cheng, X. Wang, J. Lei, D. Crandall, M. Bansal, G. Bertasius, Vindlu: a recipe for effective video-and-language pretraining, in: CVPR, 2023, pp. 10739–10750.
[9] M. Cheng, Y. Lin, M. Li, Sequence-to-sequence neural diarization with automatic speaker detection and representation, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 2719–2734.
[10] M. Cheng, et al., Computer-aided autism spectrum disorder diagnosis with behavior signal processing, IEEE Trans. Affect. Comput. 14 (2023) 2982–3000.
[11] S. Cho, M.Y. Liberman, N. Ryant, M. Cola, R.T. Schultz, J. Parish-Morris, Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations, in: Proc. Annu. Conf. Int. Speech. Commun. Assoc. (INTERSPEECH), 2019, pp. 2513–2517.
[12] S. Deng, et al., Hear me, see me, understand me: audio-visual autism behavior recognition, IEEE Trans. Multimedia 27 (2024) 2335–2346.
[13] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Conf. N. Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT), 2019, pp. 4171–4186.
[14] A. Dosovitskiy, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
[15] A. Dubey, et al., The LLaMA 3 herd of models, 2024, https://doi.org/10.48550/arXiv.2407.21783
[16] J.H. Elder, C.M. Kreider, S.N. Brasher, M. Ansell, Clinical impact of early diagnosis of autism on the prognosis and parent–child relationships, Psychol. Res. Behav. Manag. (2017) 283–292.
[17] X. Geng, et al., OSUM: advancing open speech understanding models with limited resources in academia, arXiv preprint arXiv:2501.13306, 2025.
[18] M. Ghorbian, M. Ghobaei-Arani, Large language models for mental health diagnosis and treatment: a survey, Artif. Intell. Rev. 59 (2025) 9.

[19] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: a graph convolutional neural network for emotion recognition in conversation, in: Conf. Empir. Methods Nat. Lang. Process. Int. Jt. Conf. Nat. Lang. Process, Proc. Conf. (EMNLP-IJCNLP), 2019, pp. 154–164.
[20] K. Gotham, S. Risi, A. Pickles, C. Lord, The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity, J. Autism Dev. Disord. 37 (2007) 613–627.
[21] D. Guo, et al., DeepSeek-r1: incentivizing reasoning capability in LLMs via reinforcement learning, 2025, https://doi.org/10.48550/ARXIV.2501.12948
[22] J. He, P. Li, G. Liu, S. Zhong, Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding, in: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), 2025, pp. 1–5.
[23] T. Hirota, B.H. King, Autism spectrum disorder: a review, Jama 329 (2023) 157–168.
[24] E.J. Hu, et al., LoRA: low-rank adaptation of large language models, in: ICLR, 2022.
[25] S. Jiang, T. Zheng, Y. Zhang, Y. Jin, L. Yuan, Z. Liu, Med-moe: mixture of domain-specific experts for lightweight medical vision-language models, in: Find. (EMNLP), 2024, pp. 3843–3860.
[26] D.Y. Kim, et al., Automated AI based identification of autism spectrum disorder from home videos, PJ Digit. Med. 8 (2025) 607.
[27] K.A. Kinfu, et al., Computerized assessment of motor imitation for distinguishing autism in video (CAMI-2DNet), IEEE Trans. Biomed. Eng. (2025) 1–11.
[28] A. Laguna, S. Pusil, A.L. Paltrinieri, S. Orlandi, Automatic cry analysis: deep learning for screening of autism spectrum disorder in early childhood, J. Autism Dev. Disord (2025) 1–7.
[29] C. Li, et al., Llava-med: training a large language-and-vision assistant for biomedicine in one day, Adv. Neural Inf. Process. Syst. 36 (2023) 28541–28564.
[30] J. Li, et al., Appearance-based gaze estimation for ASD diagnosis, IEEE T. Cybern. 52 (2022) 6504–6517.
[31] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, Y. Qiao, Uniformerv2: spatiotemporal learning by arming image vits with video uniformer, arXiv preprint arXiv:2211.09552, 2022.
[32] K. Li, et al., Unmasked teacher: towards training-efficient video foundation models, in: ICCV, 2023, pp. 19948–19960.
[33] K. Li, et al., Mvbench: a comprehensive multi-modal video understanding benchmark, in: CVPR, 2024, pp. 22195–22206.
[34] M. Li, et al., An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder, Comput. Speech Lang 56 (2019) 80–94.
[35] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2023) 34892–34916.
[36] W. Liu, M. Li, L. Yi, Identifying children with autism spectrum disorder based on their face processing abnormality: a machine learning framework, Autism Res. 9 (2016) 888–898.
[37] W. Liu, Y. Pan, D. Zhang, H. Deng, X. Zou, M. Li, Detecting children with autism spectrum disorder based on script-centric behavior understanding with emotional enhancement, Neurocomputing 670 (2026) 132434.
[38] W. Liu, T. Zhou, C. Zhang, X. Zou, M. Li, Response to name: a dataset and a multimodal machine learning framework towards autism study, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2017, pp. 178–183.
[39] W. Liu, et al., Self-guided DMT: exploring a novel paradigm of dance movement therapy in mixed reality for children with ASD, IEEE Trans. Vis 30 (2024) 2119–2128.
[40] Y. Liu, et al., Re-imagining multimodal instruction tuning: a representation view, in: Int. Conf. Learn. Represent. (ICLR), 2025, pp. 88901–88924.
[41] G.V. Londhe, P. Sukumar, B. Krishnan, B.Y. Supriya, R. Karthik, C. Hemalatha, Diagnosis of autism spectrum disorder using deep learning with natural language processing, in: Challenges in Information, Communication and Computing Technology, CRC Press, 2024, pp. 402–406.
[42] C. Lord, R. Luyster, Early diagnosis of children with autism spectrum disorders, Clin. Neurosci. Res. 6 (2006) 189–194.
[43] C. Lord, et al., The autism diagnostic observation schedule—generic: a standard measure of social and communication deficits associated with the spectrum of autism, J. Autism Dev. Disord. 30 (2000) 205–223.
[44] C. Lord, et al., The lancet commission on the future of care and clinical research in autism, Lancet 399 (2022) 271–334.
[45] M.J. Maenner, Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2020, MMWR Surv. Summ. 72 (2023).
[46] A. Mohanta, V.K. Mittal, Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features, Comput. Speech Lang. 72 (2022) 101287.
[47] S. Narayanan, P.G. Georgiou, Behavioral signal processing: deriving human behavioral informatics from speech and language, Proc. IEEE 101 (2013) 1203–1233.
[48] S. Natraj, N. Kojovic, T. Maillart, M. Schaer, Video-audio neural network ensemble for comprehensive screening of autism spectrum disorder in young children, PLoS One 19 (2024) e0308388.
[49] C. Okoye, et al., Early diagnosis of autism spectrum disorder: a review and analysis of the risks and benefits, Cureus 15 (2023).
[50] J.A. Omiye, H. Gui, S.J. Rezaei, J. Zou, R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, Ann. Intern. Med. 177 (2024) 210–220.
[51] I. Oosterling, et al., Improved diagnostic validity of the ADOS revised algorithms: a replication study in an independent sample, J. Autism Dev. Disord. 40 (2010) 689–703.
[52] Y. Pan, K. Cai, M. Cheng, X. Zou, M. Li, Responsive social smile: a machine learning based multimodal behavior assessment framework towards early stage autism screening, in: ICPR, 2021, pp. 2240–2247.
[53] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: ICML, 2023, pp. 28492–28518.
[54] M. Serna-Aguilera, et al., Video-based autism detection with deep learning, in: GTC, 2024, pp. 159–161.
[55] Y. Shen, Z. Xu, Q. Wang, Y. Cheng, W. Yin, L. Huang, Multimodal instruction tuning with conditional mixture of LoRA, in: Proc. Annu. Meet. Assoc. Comput Linguist. (ACL), 2024, pp. 637–648.
[56] G. Sun, et al., video-salmonn: speech-enhanced audio-visual large language models, arXiv preprint arXiv:2406.15704, 2024.
[57] C. Tang, et al., Automatic identification of high-risk autism spectrum disorder: a feasibility study using video and audio data under the still-face paradigm, EEE Trans. Neural Syst. Rehabil. Eng. 28 (2020) 2401–2410.
[58] C.K. Themistocleous, M. Andreou, E. Peristeri, Autism detection in children: integrating machine learning and natural language processing in narrative analysis, Behav. Sci. 14 (2024) 459.
[59] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nat. Med. 29 (2023) 1930–1940.
[60] Y.H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: ACL, 2019, pp. 6558.
[61] T. Wang, et al., M2pt: multimodal prompt tuning for zero-shot instruction learning, in: Conf. Empir. Methods Nat. Lang. Process, Proc. Ind. Track (EMNLP), 2024, pp. 3723–3740.
[62] P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, M.A. Armin, Vision-based activity recognition in children with autism-related behaviors, Heliyon 9 (2023) 16763.
[63] C. Xia, H. Chen, J. Han, D. Zhang, K. Li, Identifying children with autism spectrum disorder via transformer-based representation learning from dynamic facial cues, IEEE Trans. Affect. Comput. 16 (2025) 83–97.
[64] H. Xiao, et al., A comprehensive survey of large language models and multimodal large language models in medicine, Inf. Fusion 117 (2025) 102888.
[65] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, et al., Qwen3-omni technical report, arXiv preprint arXiv:2509.17765, 2025.
[66] L. Yan, C. Han, Z. Xu, D. Liu, Q. Wang, Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning, in: Int. Joint Conf. Artif. Intel. (IJCAI), 2023, pp. 1622–1630.
[67] L. Yan, Q. Wang, S. Ma, J. Wang, C. Yu, Solve the puzzle of instance segmentation in videos: a weakly supervised framework with spatio-temporal collaboration, IEEE Trans. Circuits Syst. Video Technol. 33 (2022) 393–406.
[68] L. Yan, et al., Video captioning using global-local representation, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 6642–6656.
[69] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388, 2025.
[70] J. Ye, H. Tang, Multimodal large language models for medicine: a comprehensive survey, arXiv preprint arXiv:2504.21051, 2025.
[71] S. Yin, et al., Machine learning-based early prediction model for autism spectrum disorder in infants using acoustic feature, Autism Res. 19 (3) (2026) e70179.
[72] H. Zhang, X. Li, L. Bing, Video-llama: an instruction-tuned audio-visual language model for video understanding, arXiv preprint arXiv:2306.02858, 2023.
[73] N. Zhang, M. Ruan, S. Wang, L. Paul, X. Li, Discriminative few shot learning of facial dynamics in interview videos for autism trait classification, IEEE Trans. Affect. Comput. 14 (2023) 1110–1124.

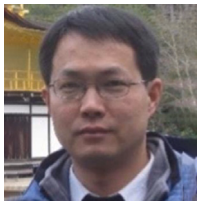
## Author biography



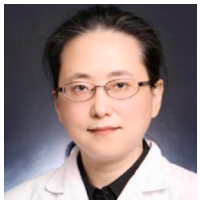
**Wenxing Liu** received the master's degree in computer science from the Chongqing University of Technology. He is currently working toward the PhD degree in computer science at Wuhan University. His research interests include ASD diagnosis, ASD assessment, and multimodal large language models.



**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 at Wuhan University. Her research interests include applications of multimodal behavior analysis to assist 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 of 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 at the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. He has extensive clinical experience in the field of developmental behavioral disorders in 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 at School of Artificial Intelligence at the Chinese University of Hong Kong, Shenzhen. He is also an Adjunct Professor at School of Artificial Intelligence at 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, and as an 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, 2024, 2025 and 2026 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 and 2025 AVSE Challenge. He received the IBM Faculty Award in 2016, the ISCA Computer Speech and Language 5-Year Best Journal Paper Award in 2018 and the Youth Achievement Award for Outstanding Scientific Research Achievements in Chinese Higher Education in 2020.