

## Hear Me, See Me, Understand Me: Audio-Visual Autism Behavior Recognition

Authors	Deng, Shijian;Kosloski, Erin E.;Patel, Siddhi;Barnett, Zeke A.;Nan, Yiyang;Kaplan, Alexander;Doan, William T.;Wang, Matthew;Singh, Harsh;Rollins, Pamela R.;Tian, Yapeng
Citation	S. Deng et al., "Hear Me, See Me, Understand Me: Audio-Visual Autism Behavior Recognition," in IEEE Transactions on Multimedia, vol. 27, pp. 2335-2346, 2025, doi: 10.1109/TMM.2024.3521838
DOI	<a href="https://doi.org/10.1109/tmm.2024.3521838">10.1109/tmm.2024.3521838</a>
Publisher	IEEE
Download date	2026-03-05 20:29:27
Link to Item	<a href="https://hdl.handle.net/20.500.14634/842">https://hdl.handle.net/20.500.14634/842</a>

# Hear Me, See Me, Understand Me: Audio-Visual Autism Behavior Recognition

Shijian Deng , Erin E. Kosloski , Siddhi Patel , Zeke A. Barnett, Yiyang Nan , Alexander Kaplan, Sisira Aarukapalli, William T. Doan , Matthew Wang, Harsh Singh, Pamela R. Rollins , and Yapeng Tian , *Member, IEEE*

**Abstract**—In this article, we introduce a novel problem of audio-visual autism behavior recognition, which includes social behavior recognition, an essential aspect previously omitted in AI-assisted autism screening research. We define the task at hand as one that is audio-visual autism behavior recognition, which uses audio and visual cues, including any speech present in the audio, to recognize autism-related behaviors. To facilitate this new research direction, we collected an audio-visual autism spectrum dataset (AV-ASD), currently the largest video dataset for autism screening using a behavioral approach. It covers an extensive range of autism-associated behaviors, including those related to social communication and interaction. To pave the way for further research on this new problem, we intensively explored leveraging foundation models and multimodal large language models across different modalities. Our experiments on the AV-ASD dataset demonstrate that integrating audio (mainly ambient sound), visual, and speech (predominately spoken language) modalities significantly enhances the performance in autism behavior recognition. Additionally, we explored the use of a *post-hoc to ad-hoc* pipeline in a multimodal large language model to investigate its potential to augment the model’s explanatory capability during autism behavior recognition.

**Index Terms**—Audio, autism, benchmark, dataset, explainability, multimodal large language model, speech, video.

Received 22 March 2024; revised 31 July 2024; accepted 1 September 2024. Date of publication 23 December 2024; date of current version 12 May 2025. This work was supported by the UTD SPIRe grant. The associate editor coordinating the review of this article and approving it for publication was Prof. Bingkun Bao. (*Corresponding author: Yapeng Tian.*)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by UTD Institutional Review Board under Application No. IRB-23-622, and performed in line with the A comprehensive audio-visual autism related behavior dataset.

Shijian Deng, Alexander Kaplan, Sisira Aarukapalli, William T. Doan, Matthew Wang, and Yapeng Tian are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: shijian.deng@utdallas.edu; sasha.kaplan@utdallas.edu; sisira.aarukapalli@utdallas.edu; william.doan@utdallas.edu; matthew.wang@utdallas.edu; yapeng.tian@utdallas.edu).

Erin E. Kosloski, Siddhi Patel, and Pamela R. Rollins are with the School of Behavioral and Brain Sciences, The University of Texas at Dallas, Dallas, TX 75235 USA (e-mail: erin.kosloski@utdallas.edu; siddhi.patel@utdallas.edu; rollins@utdallas.edu).

Zeke A. Barnett is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: zbarnett@andrew.cmu.edu).

Yiyang Nan is with the Department of Computer Science, Brown University, Providence, RI 02912 USA (e-mail: yiyang\_nan@brown.edu).

Harsh Singh is with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi 23201, UAE (e-mail: harsh.singh@mbzuai.ac.ae).

Our dataset, code, and models are available at: <https://github.com/ShijianDeng/AV-ASD>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TMM.2024.3521838>, provided by the authors.

Digital Object Identifier 10.1109/TMM.2024.3521838

## I. INTRODUCTION

AUTISM spectrum disorder (ASD) is a complex, heterogeneous neurodevelopmental condition associated with persistent challenges in social communication and interaction, as well as the presence of restrictive or repetitive behaviors and interests [1]. The prevalence of autism has steadily risen in recent decades. Early identification and intervention are critical to support autistic<sup>1</sup> children in developing social communication, language, and adaptive functioning [3], [4]. However, the worldwide average age at which ASD is diagnosed is 60.48 months [5], as the diagnosis process is time-consuming due to the necessity of long-term behavioral observation and the limited availability of specialists. Current early screening efforts primarily use parent report instruments (e.g., [6]) which are suitable only for children under 30 months and are inherently subjective.

Hence, there is a critical demand for an efficient, accessible, and objective screening tool that can effectively be used for a broader age range of children. To this end, researchers are exploring the application of artificial intelligence (AI) to autism screening. Previous studies have investigated the application of AI to autism detection by using various objective data-driven approaches such as functional Magnetic Resonance Imaging (fMRI) scans, eye tracking data, and behavioral observations (e.g., [7], [8], [9]). In our current study, we employ the behavioral observation method to screen autism, using the clinical gold standard criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), i.e. This involves analyzing both social interaction challenges and restricted and repetitive behaviors (RRBs) from video datasets [1]. Prior research that has utilized behavioral observation methods by using video datasets to identify autistic behaviors has focused solely on the visual modality. This singular focus has limited the scope of analyses as it tends to capture only RRBs overlooking social interaction challenges integral to the diagnosis of autism. Further, RRBs are not always present before 36 months, whereas social challenges are detectable in the first year of life [10], [11]. Detecting social interaction difficulties, however often requires moving beyond the visual modality and accessing auditory information. Thus it is imperative for an AI early autism screening tool to be capable of identifying behaviors related to autistic individuals’ social interaction difficulties. Therefore, in an effort to better

<sup>1</sup>We use identity-first (i.e., autistic person) rather than person-first (i.e., person with autism) as preferred by most autistic self-advocates [2].

leverage AI for screening autism we created a new audio-visual autism behavior recognition dataset (AV-ASD) and introduced an audio-visual autism behavior recognition task, which aims to identify both social interaction behaviors in addition to RRBs. The AV-ASD dataset, featuring 928 video clips from 569 unique videos across 10 categories, is currently the largest for autism screening using a behavioral approach.

With the newly collected dataset, we establish a comprehensive benchmark for exploring how to better recognize autism behaviors in videos. We develop several baselines and novel frameworks using strong foundation models like CLIP (image), ImageBind (video/audio), and Whisper (speech). We further investigate the multimodal integration with temporal modeling and evaluate the effectiveness of Multimodal Large Language Models (MLLMs), including GPT-4o [12] and LLaVA [13], as zero-shot benchmarks. To utilize audio and speech cues in MLLMs, we adopt audio captioning and speech recognition models to generate text prompts. To further improve performance, we employ an audio-visual instruction tuning, adapting LLaVA into LLaVA-ASD with our annotated data. This significantly enhances its efficacy, particularly with audio-augmented prompts. However, solely relying on behavior labels during tuning can compromise the model's explainability and lead to catastrophic forgetting. To address these challenges, we propose a novel *post-hoc to ad-hoc* framework that maintains the model's predictive accuracy while preserving its prediction explanation ability.

Our contributions are as follows:

- 1) We introduce a comprehensive autism behavior recognition dataset containing 928 clips across 10 categories, annotated from 569 videos. This is currently the largest dataset for autism screening in unconstrained videos. The dataset includes both restricted, repetitive patterns of behavior and unique social interaction behaviors relevant to autism, which were omitted in previous similar datasets.
- 2) We extensively benchmark different state-of-the-art foundation models with diverse modality combinations and discover that specialized MLLMs, integrating audio-visual cues, are capable of tackling autism behavior recognition.
- 3) Our *post-hoc to ad-hoc* framework demonstrates potential in mitigating catastrophic forgetting during instruction tuning, while also achieving explainable predictions.

## II. RELATED WORK

### A. AI and Datasets in ASD Research

Datasets play a crucial role in AI-powered autism detection. Traditionally, these datasets are comprised of videos recorded exclusively in a laboratory setting [14], [15], [16], [17], [18], [19], [20], [21], [22]. However, this approach inherently limits the datasets to controlled environments and scenarios. There have been commendable strides towards introducing more realistic, unbounded datasets such as the Self-Stimulatory Behavior Dataset (SSBD) [23], the Expanded Stereotype Behavior Dataset (ESBD) [24], the dataset curated by Wei et al. [25], and the Autism Stimming Behavior Dataset (ASBD) [26]. Despite

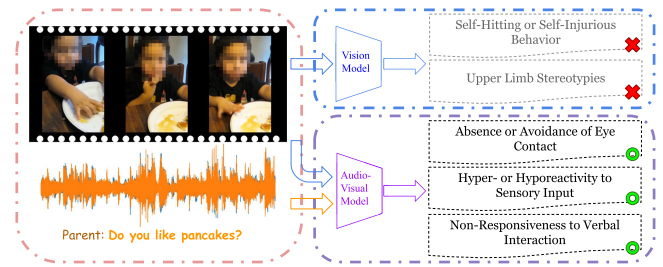


Fig. 1. The vision-only model incorrectly identified two behaviors that were not present, whereas the audio-visual model correctly identified the three behaviors present in the clip. This illustrates how multimodal integration enables more accurate behavior identification.

TABLE I  
COMPARISON OF THE AV-ASD DATASET WITH OTHER AUTISM-RELATED BEHAVIOR DATASETS USED IN ASD SCREENING RESEARCH

Dataset	Clips	Categories	Multi-Label	Social Behaviors
SSBD [23]	75	3	✗	✗
ESBD [24]	141	4	✗	✗
Wei et al. [25]	61	3	✗	✗
ASBD [26]	165	4	✗	✗
AV-ASD	<b>928</b>	<b>10</b>	✓	✓

this progress, these datasets primarily focus on RRBs, neglecting the atypical social interactions necessary for autism diagnosis. Leveraging these RRBs datasets, researchers have proposed various vision-based models for early ASD detection [23], [24], [25], [27], [28], [29]. However, relying solely on vision overlooks the valuable social information that can be captured only with audio. Addressing these limitations, our work introduces a new multimodal ASD dataset and a novel multimodal learning framework to best leverage audio-visual cues.

### B. Audio-Visual Learning

Audio-visual learning has shown its potential in various applications, including audio-visual action, emotion or speech recognition [30], [31], [32], [33], [34], [35], audio-visual localization [36], [37], [38], [39], [40], [41], [42], [43], [44], audio-visual video parsing [45], [46], [47], audio-visual segmentation [48], [49], [50], audio-visual source separation [51], [52], [53], audio-visual question answering or dialogue [54], [55], [56], [57], and audio-visual large language model [58], [59], [60]. The concurrent analysis of audio and visual data can provide a more holistic understanding of complex behaviors, particularly those in which the audio component plays a critical role. In this work, we have integrated audio data into our model, demonstrating that audio-visual learning can substantially enhance the efficiency of identifying a range of autistic behaviors.

## III. THE AV-ASD DATASET

The Audio-Visual Autism Spectrum Dataset (AV-ASD) is a curated collection designed to enhance research into autism-related behaviors, especially in social contexts. It includes 928 clips from 569 YouTube and Facebook videos, capturing

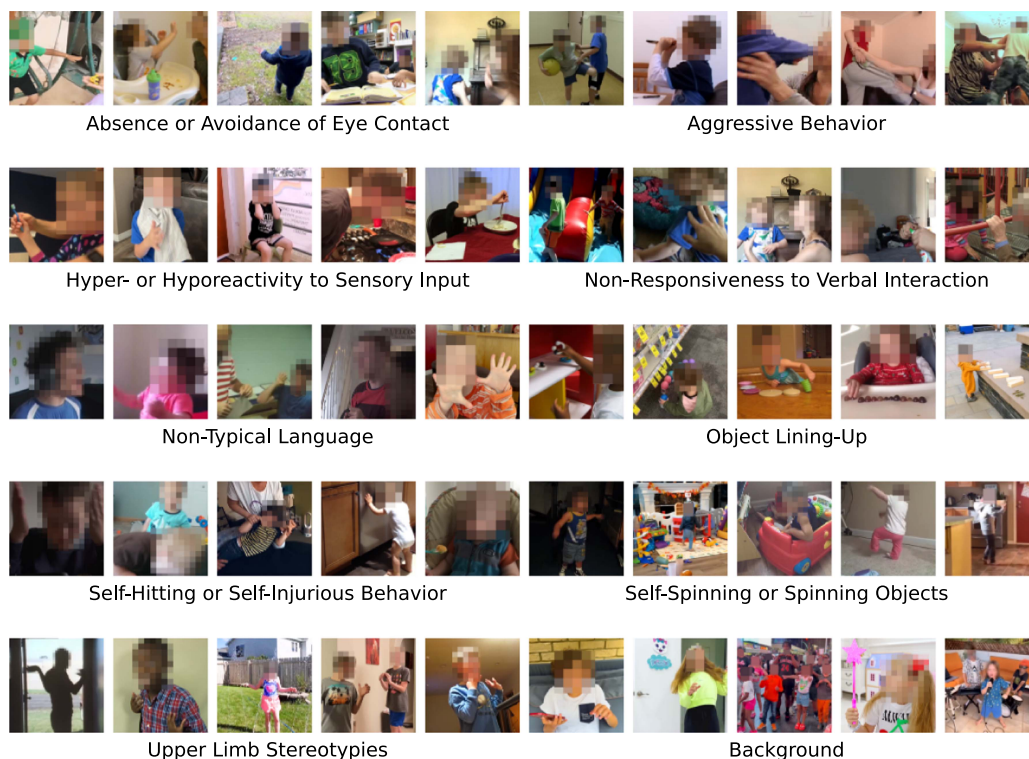


Fig. 2. A depiction of the AV-ASD dataset, illustrating five sample instances from each category.

diverse behaviors and environments. Some examples are shown in Fig. 2. AV-ASD distinguishes itself from preceding datasets in several significant ways, as delineated in Table I. First, our dataset offers a far greater number of categories and video clips than all previous datasets combined. Second, AV-ASD is the first dataset to include social behaviors. Third, our dataset is the first ever autism dataset to use a multi-label setting, which is more practical since multiple autism-related behaviors could happen at the same time. Lastly, each instance in the AV-ASD dataset is meticulously annotated with time-stamped labels that identify various autistic behaviors. This stands in contrast to some previous datasets, such as ESB, which are weakly labeled and lack specific start and end times for the observed behaviors. These make AV-ASD a pioneering dataset for AI-based multimodal autism research.

#### A. Multimodal Nature of Social Behaviors

Detection and interpretation of social behaviors often require both auditory and visual information. In particular, the context of social interactions is deeply intertwined with the content of conversations. As such, audio, specifically speech, offers an essential layer of contextual data invaluable for social behavior analysis. For example, as shown in Fig. 3, a child answering another person's question could be visually interpreted as a normal communicative engagement. Yet, when the speech content is analyzed, it is clear that the child's response bears no relevance to the posed question, illustrating the critical role of multimodal analysis.

Thus, our multimodal framework aids in constructing a more accurate model of social autistic behaviors.



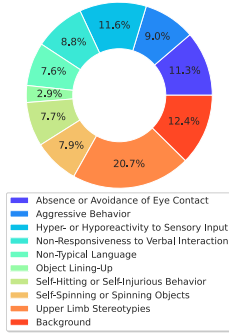
Fig. 3. An ASD child responds "I have hands." to the person who asks "How old are you?".

#### B. Data Collection, Annotation, and Statistics

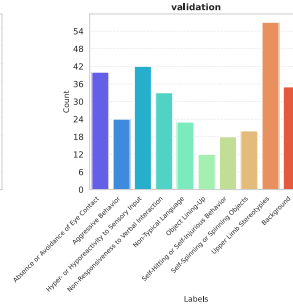
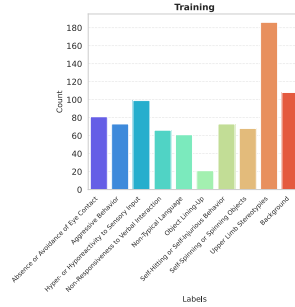
To create our behavioral categories for the social and RRBs domains of autism, we identified social challenges from the social behavior classifications of the DSM-V-TR [1] and M-CHAT-R/F [6] screening tool. RRBs were adapted from SSBD [23] and ESB [24]. Notably, we were limited to behaviors that could be identified in a brief video clip. The resulting taxonomy consisted of nine distinct autistic behavioral categories and one *Background* (i.e., not-applicable) category.

We curated the AV-ASD dataset through a keyword video search, excluding irrelevant content such as lectures and cartoons. This process resulted in 928 distinct video clips extracted from 569 online videos. A team of six volunteer students meticulously annotated each clip, which was then verified by a Speech-Language Pathologist (SLP) with 15 years of experience working with autistic children.

Our final AV-ASD dataset consists of 928 clips in 10 categories, amounting to roughly 6 hours and 40 minutes of footage,



(a) Category-wise distribution.



(b) Distribution of video in different splits.

Fig. 4. Statistical illustrations of the AV-ASD dataset.

extracted from 569 videos with a total duration exceeding 86 hours. A detailed statistical analysis of our dataset is provided in Fig. 4. For each category: *Absence or Avoidance of Eye Contact* includes 161 clips; *Aggressive Behavior* comprises 129 clips; *Hyper- or Hyporeactivity to Sensory Input* incorporates 166 clips; *Non-Responsiveness to Verbal Interaction* contains 126 clips; *Non-Typical Language* has 108 clips. *Object Lining-Up* incorporates 42 clips; *Self-Hitting or Self-Injurious Behavior* comprises 110 clips; *Self-Spinning or Spinning Objects* includes 113 clips; *Upper Limb Stereotypies* consists of 296 clips; and We also added a *Background* category that contains 177 clips. All clips were thoroughly reviewed for their relevance and validity in relation to our study. We randomly divided the dataset into a training set (553 clips), a validation set (193 clips), and a testing set (182 clips). Clips have a mean duration of 25.88 s, median duration of 10.00 s, maximum duration of 887.01 s, and minimum duration of 1.00 s. Each clip has an average of 1.54 behavior categories.

Our dataset covers a very diverse range of ages from babies to teenagers, which allows tools built based on it to be used for a broader age range of children.

### C. Annotators and Instructions for Annotation

All student annotators are residents of either the United States, China, or India. Out of the 6 students, 5 are male: 3 are from the United States, including one female, 2 from China, and 1 from India. The SLP is a female from the United States.

To acquire annotations, we implemented a two-step process:

- 1) We instructed student annotators to review a given video and check for any signs of autistic behavior. If present, they were to label it in a single-label setting with the most significant category and record the start and end times of the behavior.
- 2) We extracted clips from the video for fine-grained labeling. The SLP's task was to re-label the clips in a multi-label setting, confirming whether each autistic behavior appeared in each clip, without the need to record the start and end times again.

## IV. AUTISM BEHAVIOR RECOGNITION

Building on recent advances in large models, we develop a novel set of methods exploiting both audio and visual

modalities to identify autism-related behaviors in videos. These models are benchmarked on the AV-ASD dataset to evaluate their performance and establish baselines for future research.

We leverage recent foundation models to extract features for autism behavior recognition, utilizing CLIP [61] for image features, ImageBind [62] for processing images, videos, and audio, and Whisper [63] for speech analysis. These extracted features are then inputted into either linear probes, Multi-Layer Perceptrons (MLPs), or temporal models for prediction.

*Image Representation:* To utilize CLIP and ImageBind as image encoders, a video is transformed into a single composite image. Specifically, given a sequence of frames  $V = \{F_1, F_2, \dots, F_n\}$ ,  $p \times q$  frames are uniformly selected and arranged into a  $p \times q$  grid to form a composite image denoted as  $I_V$ . CLIP or ImageBind then extracts features from this image, represented by  $f_{\text{CLIP}}(I_V)$  for CLIP and  $f_{\text{ImageBind}}(I_V)$  for ImageBind, respectively. These extracted features are subsequently utilized in logistic regression or MLP models designed to classify various autism behaviors. Additionally, this composite image format serves as instrumental visual input for MLLMs.

*Audio and Video Representation:* For both audio and video, we employ clip-level and segment-level representations. In the segment-level approach, an input video with accompanying audio is first divided into  $T$  non-overlapping pairs of visual and audio segments, denoted as  $\{V_t, A_t\}_{t=1}^T$ , with each segment lasting 1 s. During the training phase, each segment is associated with a behavioral label  $y_t$ . The audio feature vector for each audio segment  $A_t$  is encoded by  $f_{a_t} = \text{ImageBind}(A_t)$ , and similarly, the video feature vector for each video segment  $V_t$  is encoded by  $f_{v_t} = \text{ImageBind}(V_t)$ , utilizing the pre-trained ImageBind model.

*Speech Representation:* To encode each audio segment  $A_t$  into a speech feature vector  $f_{s_t} = \text{pool}(\text{Whisper}(A_t))$ , we utilize the encoder of the Whisper model [63], applying an average pooling over the time dimension at the end.

*Temporal Modeling:* For each modality  $m$  (e.g., audio, visual, speech), we process the segment-wise features for temporal modeling. These features are aggregated over adjacent ten segments to form feature sequences, represented as  $T_{t \rightarrow t+9}^m = T(F_{t \rightarrow t+9})$ . Subsequently, a transformer encoder layer is applied to the sequence of each modality. This approach effectively integrates temporal information, resulting in a temporally aggregated feature representation denoted by  $x_m^T$ . Along with

features extracted directly from the entire clip without temporal modeling, all features are collectively represented as  $x_m$ .

**Multimodal Fusion:** To integrate information across modalities, we leverage several multimodal fusion methods: average, max, concatenation, and weighted fusion.

- **Average Fusion:** The fused feature vector  $f$  is computed as the average of features  $x_m$  across all modalities  $m$ . This is denoted as:

$$f = \frac{1}{M} \sum_{m=1}^M x_m, \quad (1)$$

where  $M$  represents the total number of modalities.

- **Max Fusion:** The fused feature vector  $f$  is computed as the maximum feature  $x_m$  across all modalities  $m$ :

$$f = \max_{m=1}^M x_m. \quad (2)$$

- **Concatenation Fusion:** The fused feature vector  $f$  is computed by concatenating features  $x_m$  from all modalities  $m$ . This is denoted as:

$$f = [x_1, x_2, \dots, x_M]. \quad (3)$$

- **Weighted Fusion:** The fused feature vector  $f$  is computed as the weighted average of features  $x_m$  across all modalities  $m$  with weights  $w_m$ . This is denoted as:

$$f = \sum_{m=1}^M w_m \cdot x_m, \quad (4)$$

where  $M$  represents the total number of modalities.

### A. Zero-Shot Baselines With MLLMs

MLLMs have revolutionized zero-shot learning, seamlessly integrating information across modalities like vision and language. To this end, we investigated the potential of MLLMs in precisely identifying autism behaviors in videos, focusing on their zero-shot capabilities. We employed two MLLMs: GPT-4o, a state-of-the-art proprietary model developed by OpenAI [12], and LLaVA, an open-source alternative excelling in similar tasks [64], [65] as the benchmark.

**MLLMs for Autism Behavior Recognition:** Given that current open-source SOTA MLLMs lack the ability to process long video sequences, we opted to repurpose the composite image  $I_V$  as the visual input for our MLLMs. This approach has also demonstrated competitive performance in [66]. To assist the MLLMs in accurately identifying autism-related behaviors, we developed a textual prompt  $P$  (see Fig. 5). This prompt is strategically devised to act as a linguistic guide, steering the MLLMs toward recognizing autistic cues and patterns inherent in the video content. The prediction of behavior is thus derived using the formula  $\hat{y} = \text{MLLM}(I_V, P)$ , where  $\hat{y}$  represents the MLLMs' output.

**Bridging the Multimodal Gap:** Current MLLMs typically focus on image and text inputs, posing a challenge for analyzing multimodal data including audio and speech cues. To overcome

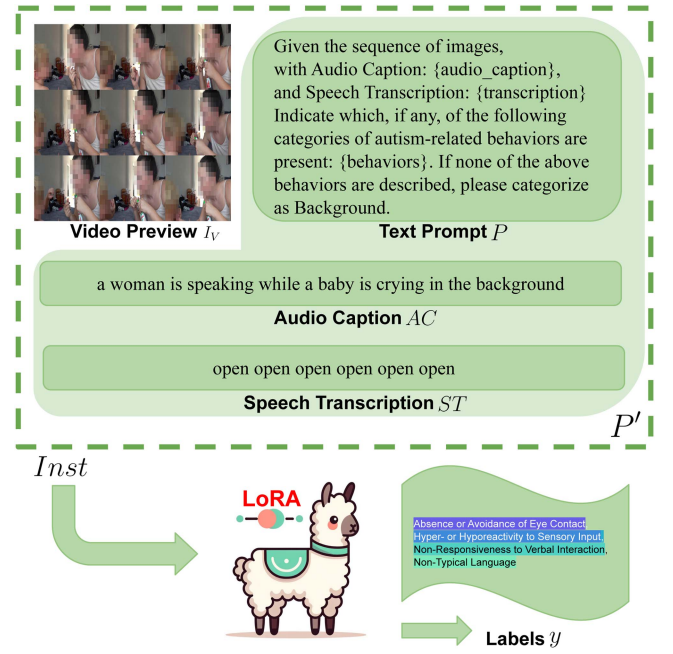


Fig. 5. LLaVA-ASD: Instruction Tuning for LLaVA. Given a video preview  $I_V$  and an enhanced text prompt  $P'$ , which is a text prompt  $P$  augmented with an audio caption  $AC$  and speech transcription  $ST$ . These elements are combined to form the model's instruction input  $Inst$ . The output consists of multiple autism behavior labels presented in text format as  $y$ . We employed LoRA for efficient fine-tuning.

this limitation in our autism behavior recognition task, we propose a Multimodal Representational Text Fusion with two key strategies:

- 1) **Audio Captioning [67]:** Converts audio segments into textual descriptions, adding information from the audio to the input.
- 2) **Speech Recognition [63]:** Transcribes spoken segments into text, giving the model direct linguistic cues from the audio.

By leveraging these strategies, we translate audio information into structured text representations readily usable by MLLMs. Combining this textual data with the prompt  $P$  enables comprehensive multimodal analysis, empowering MLLMs to capture multimodal cues in video content, ultimately leading to a more accurate recognition of autism-related behaviors.

### B. MLLMs With Instruction Tuning

Beyond zero-shot testing, we leverage the power of instruction tuning [58], [64], [68], [69] to enhance MLLMs' performance on our specialized dataset of video and audio data containing autism-related behaviors. This aims to refine the models' understanding of autism-specific cues from different modalities, leading to heightened effectiveness in identifying autism-related behaviors. Since GPT-4o is not open-source, we adopt LLaVA as a baseline for the instruction tuning study.

For training, we construct an instruction tuning pair denoted as  $\{Inst, y\}$ . Here,  $Inst = [I_V, P']$  represents the model's input for a video clip, where  $P'$  is an enhanced text prompt. This

TABLE II  
AUTISM BEHAVIOR RECOGNITION RESULTS OF DIFFERENT BASELINES ON THE AV-ASD TEST SET

Method	F1-score (%)
Dummy Baseline	26.83
CLIP (ViT-L/14@336px, linear probe)	45.03
CLIP (ViT-L/14@336px, MLP)	45.72
ImageBind (image, linear probe)	12.12
ImageBind (image, MLP)	39.19
ImageBind (video, linear probe)	10.14
ImageBind (video, MLP)	44.06
ImageBind (video, temporal)	50.87
ImageBind (audio, linear probe)	27.54
ImageBind (audio, MLP)	28.47
ImageBind (audio, temporal)	41.74
Whisper (speech, linear probe)	34.04
Whisper (speech, MLP)	36.48
Whisper (speech, temporal)	39.69
GPT-4o zero-shot (vision)	43.69
GPT-4o zero-shot (vision, audio)	35.10
GPT-4o zero-shot (vision, speech)	49.58
GPT-4o zero-shot (vision, audio, speech)	43.42
LLaVA zero-shot (vision)	13.80
LLaVA zero-shot (vision, audio)	4.20
LLaVA zero-shot (vision, speech)	17.88
LLaVA zero-shot (vision, audio, speech)	12.39
ImageBind + Whisper (linear probe)	37.36
ImageBind + Whisper (MLP)	39.97
ImageBind + Whisper (temporal)	53.17
Llama3-8B	49.07
Video-LLaMA2-7B	51.20
Ours (LLaVA-ASD)	<b>61.91</b>

The **best** and **second best** results are highlighted. Llama3-8B used additional video captions generated by VAST's vision captioner. Video-LLaMA2-7B used raw audio-visual clips as input.

prompt  $P'$  is a combination of the initial input prompt  $P$  augmented with audio caption (AC) and speech transcription (ST). The term  $y$  refers to the annotated label for the behavior categories in the video clip. In implementation, we employed the low-rank adaptation (LoRA) [70] for efficient training. We denote the trained model as LLaVA-ASD.

## V. EXPERIMENTS

### A. Experimental Setup

For benchmarking purposes, we conducted a performance evaluation of several baseline models using our newly collected AV-ASD dataset. Below are the experimental details for these baselines.

*Dummy Baseline:* The dummy baseline is essentially a blind guess and doesn't require any input. Since we are using the F1 score as our metric, we used a common strategy to maximize it: assigning a positive prediction 1 to all cases. This approach helps us compare how much useful knowledge the model has learned compared to this simple baseline.

*Image Grid Creation:* For creating an image grid for the LLaVA model, we chose 5 different variants: 4 (2 × 2), 6 (3 × 2), 9 (3 × 3), 12 (4 × 3), and 16 (4 × 4), according to the settings in [66].

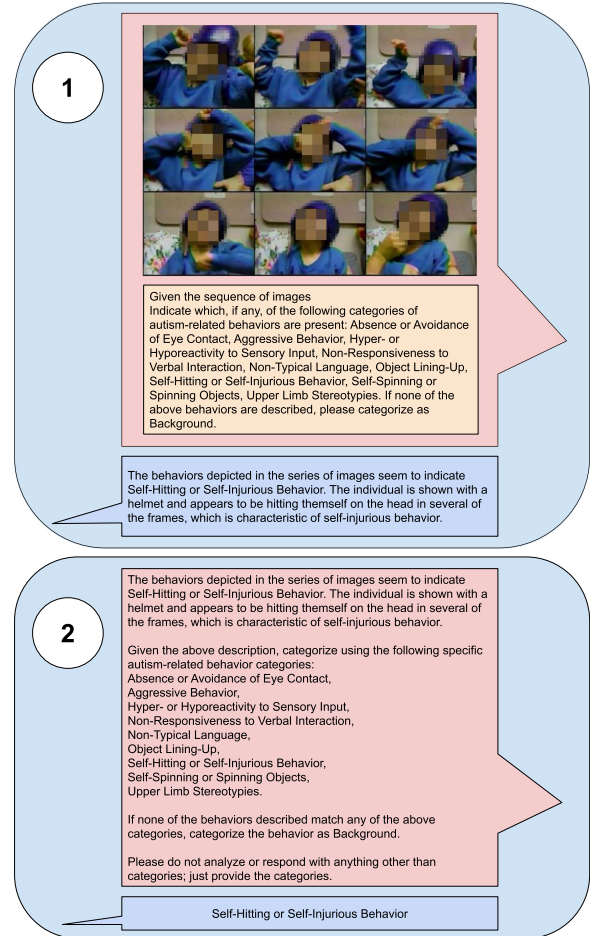


Fig. 6. The prompt we use to obtain zero-shot inference results from GPT-4o and LLaVA. Two steps are separately presented in two boxes. The red messages are user inputs and the blue messages are model responses.

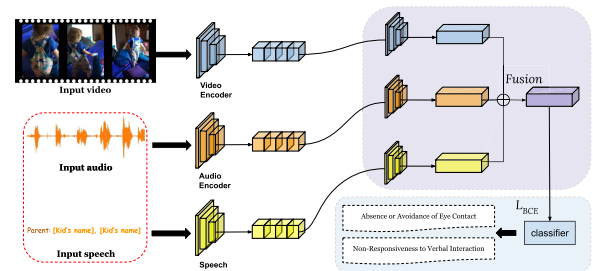


Fig. 7. An overview of our classification-based (non-LLM) framework.

*Features Extraction:* For image features, we utilized CLIP's top-performing model, ViT-L/14@336px, for visual feature extraction. For audio and video, the ImageBind huge model [62] was employed to extract a 1024-dimensional (1024-D) feature from each clip. Additionally, we used ImageBind to extract image features as well, facilitating a comparison with CLIP. For speech, Whisper's medium model [63] was used, producing a 1024-D embedding from the encoder after pooling, thus matching the feature size extracted from ImageBind. This feature size compatibility enables effective feature fusion in later stages.

TABLE III

F1-SCORE (%) ON AV-ASD AUTISM-RELATED BEHAVIORS RECOGNITION RESULTS BY OUR CLASSIFICATION-BASED MODEL (NON-LLM) WITH DIFFERENT MODALITIES AND FUSION METHODS (FOR SINGLE MODALITY, NO FUSION IS REQUIRED, BUT WE PROVIDE THE PERFORMANCE IN THE AVERAGE COLUMN FOR CONVENIENCE OF COMPARISON)

Method	Average	Max	Concat	( $a : v : s = 2 : 1 : 1$ )	( $a : v : s = 1 : 2 : 1$ )	( $a : v : s = 1 : 1 : 2$ )
Audio	41.74	N/A	N/A	N/A	N/A	N/A
Visual	50.87	N/A	N/A	N/A	N/A	N/A
Speech	39.69	N/A	N/A	N/A	N/A	N/A
Audio-Visual	51.46	52.06	52.22	N/A	N/A	N/A
Audio-Speech	42.13	42.80	42.28	N/A	N/A	N/A
Visual-Speech	51.63	51.40	52.22	N/A	N/A	N/A
Audio-Visual-Speech	<u>52.70</u>	51.76	51.96	50.83	52.56	<b>53.17</b>

The last three columns represent weighted ratio sets that we use for the weighted fusion. The **best** and second best results are highlighted.

TABLE IV

AUTISM BEHAVIOR RECOGNITION RESULTS WITH DIFFERENT MODALITIES BY LORA FINE-TUNED ON LLaVA

Behavior	V	V+A	V+S	V+A+S
Absence or Avoidance of Eye Contact	42.86	49.32	<b>59.46</b>	<u>55.70</u>
Aggressive Behavior	<b>75.36</b>	64.00	69.70	<u>74.29</u>
Hyper- or Hyporeactivity to Sensory Input	<u>43.75</u>	39.47	40.00	<b>44.07</b>
Non-Responsiveness to Verbal Interaction	35.71	32.65	<u>40.00</u>	<b>43.08</b>
Non-Typical Language	11.43	22.22	<u>27.91</u>	<b>35.90</b>
Object Lining-Up	70.59	<b>88.89</b>	82.35	<b>88.89</b>
Self-Hitting or Self-Injurious Behavior	50.00	<u>57.14</u>	<u>57.14</u>	<b>57.89</b>
Self-Spinning or Spinning Objects	66.67	61.22	<u>67.92</u>	<b>71.43</b>
Upper Limb Stereotypies	56.31	61.02	<u>64.06</u>	<b>64.81</b>
Background	83.08	<u>87.50</u>	<b>88.89</b>	83.08
<b>Average</b>	53.58	56.34	<u>59.74</u>	<b>61.91</b>

The **best** and second best results are highlighted.

TABLE V

F1 SCORES FOR DIFFERENT FRAME COMBINATIONS

Frames	V	V+A	V+S	V+A+S
4 (2x2)	51.81	53.59	56.73	56.96
6 (3x2)	53.45	54.91	59.44	<b>61.91</b>
9 (3x3)	53.48	56.34	54.93	<u>61.75</u>
12 (4x3)	53.58	54.84	55.03	58.81
16 (4x4)	52.95	55.88	59.74	58.43

The **best** and second best results are highlighted.

We used the CoNeTTE [67], pretrained on Clotho [71], to generate audio captions (AC) and employed Whisper large-v3 [63] to generate speech transcriptions (ST) for each clip.

*Linear Probes and Non-Linear Mapping:* A logistic regression model was trained with the settings: random state set to 0 and maximum iterations limited to 3000, aiming to predict behavior categories based on each modality feature individually. We further explored multimodal fusion through four methods: average, max, concatenation, and weighted fusion. For the latter, we allowed the weights for each modality to be either 1.0 or 2.0, before applying linear probes to the fused features. Transitioning from logistic regression, we employed an MLP for nonlinear mapping using the MLPClassifier from scikit-learn, with the random state set to 0 and maximum iterations at 3000. The same fusion strategies were applied, and the outcomes were documented in Table II.

*Temporal Modeling:* Using ImageBind, we extracted a 1024-dimensional feature from a one-second audio segment and another 1024-dimensional feature from a one-second visual segment. Additionally, Whisper’s medium model was deployed to

extract a 1024-dimensional speech feature from a one-second audio segment. For temporal modeling, we focused on a time sequence of 10 seconds, concatenating features from each second to form a sequence spanning 10 seconds in length. An intensive hyperparameter search was conducted, experimenting with various numbers of nodes, layers, and transformer layer heads, as well as dropout rates, weight decay factors, etc. We utilized the Adam optimizer with an initial learning rate of 0.001 and trained all models for 100 epochs with a batch size of 928. Cross-entropy loss with class weights was employed to tackle imbalance issues. The framework is demonstrated in Figure 7.

*MLLMs:* To test the zero-shot performance of MLLMs, we used the OpenAI API to get the response of GPT-4o with the specific checkpoint of gpt-4o-2024-05-13.<sup>2</sup> For LLaVA, we used the Hugging Face checkpoint llava-v1.5-13b.<sup>3</sup>

The prompt we used to obtain inference results from GPT-4o and LLaVA is shown in Figure 6 and involves a two-step pipeline:

- 1) We utilized an image and the task description to request the model to recognize autism behaviors. However, often-times, the model adopted a freestyle approach and did not directly provide behavior labels.
- 2) We used the output from the previous model as input and sent it to GPT-4o, allowing it to reformat the answer into a simple behavior label format. This enabled us to easily evaluate performance through word matching.

For fine-tuning the LLaVA model, we enabled LoRA with  $r = 128$  and  $\alpha = 256$ . The learning rate for the multimodal projector was set at  $2 \times 10^{-5}$ . The model was trained using DeepSpeed with configurations from zero3. CLIP ViT-L/14@336px was employed as the vision encoder. Each model underwent training for 100 epochs, with inference conducted on the validation dataset every 10 epochs to identify the optimal epoch. The models’ performance on the testing dataset is documented in Table IV.

Additionally, to check and compare with the fine-tuned text-only LLM and text-free MLLM, we chose Llama3 8B [72] and Video-LLaMA2 7B [60] for our experiments. We used VAST’s vision captioner [73] to caption all the videos. By combining these video captions with audio captions and speech

<sup>2</sup>[Online]. Available: <https://openai.com/index/hello-gpt-4o/>

<sup>3</sup>[Online]. Available: <https://huggingface.co/liuhaotian/llava-v1.5-13b>



baselines listed in Table II, thereby illustrating the superior capability of MLLMs in identifying autism behaviors.

- 2) The LLaVA-ASD model, incorporating visual, audio, and speech information, achieves the best performance in many categories and the best overall performance, emphasizing the importance of multimodal data for accurate recognition of autism behaviors. This suggests that the use of audio captioning and speech transcription can leverage auditory and speech information into prompts, thus improving recognition in videos. Additionally, these results confirm that the MLLM serves as an efficient integrator of different modalities.

### C. Beyond Recognition: Explainability

For an AI assistant to effectively aid doctors in autism screening, it is essential for the system to provide explanations for its suggestions. MLLMs demonstrate strong reasoning capabilities and can generate detailed explanations alongside predictions. However, zero-shot models may yield incorrect autism behavior recognition results, as illustrated in Figure 9(b). While instruction tuning using ground truth behavior labels improves performance, this approach risks reducing the model to a mere classifier, potentially losing its reasoning ability due to catastrophic forgetting of previous knowledge (refer to Fig. 9(c)). To overcome this issue and achieve explainable predictions, we initiate an exploration, laying groundwork for further advancements in future research.

A straightforward solution to ensure accurate explanations is to utilize ground truth annotations from human experts for fine-tuning. However, this process is labor-intensive and costly. To circumvent these challenges, we propose a novel, efficient self-supervised pipeline: *post-hoc to ad-hoc*, illustrated in Fig. 8. It contains two steps. First, with a visual input, audio caption, speech transcription, and prompt instruction  $P_{\text{inst}}$ , accompanied by ground truth behavior labels  $L_{gt}$ , we employ LLaVA to infer explanations for these labels. The output of this step, termed *post-hoc reasoning*:

$$\mathbf{R}_{\text{post-hoc}} = \text{MLLM}(I_V, AC, ST, P_{\text{inst}}, L_{gt}) \quad (5)$$

Second, in the absence of ground truth, we utilize the *post-hoc reasoning* as pseudo labels to train our model for generating *ad-hoc reasoning*, aiming for outputs similar to the *post-hoc reasoning*:

$$\mathbf{R}_{\text{ad-hoc}} = \text{MLLM}(I_V, AC, ST, \hat{P}_{\text{inst}}; \theta) \quad (6)$$

$$\hat{\theta} = \arg \min_{\theta} CE(\mathbf{R}_{\text{ad-hoc}}, \mathbf{R}_{\text{post-hoc}}) \quad (7)$$

Here,  $\hat{P}_{\text{inst}}$  denotes the prompt for the *ad-hoc* step and  $CE(\cdot)$  is cross-entropy loss. Both  $\mathbf{R}_{\text{ad-hoc}}$  and  $\mathbf{R}_{\text{post-hoc}}$  are text outputs of models.

Fig. 9(d) illustrates the explanation results generated by our explainable framework. We can see that fine-tuning MLLMs with synthetic *post-hoc reasoning* data effectively prevents the model from reducing to a trivial behavior classifier. Additionally, it significantly improves its explainability for the task of recognizing autism behaviors.



Fig. 9. Explainability of different models. In this example, the ground truth labels are: *Absence or Avoidance of Eye Contact* and *Non-Responsiveness to Verbal Interaction*. We can see that the LLaVA model with zero-shot cannot even predict the correct behaviors; the one tuned directly on ground truth labels cannot explain the decision-making and has a catastrophic forgetting issue. Our model tuned with *post-hoc to ad-hoc* framework correctly predicts the behaviors and gives explanations. It also follows other instructions well.

## VI. FUTURE DIRECTIONS

Opportunities for expanding the size of the dataset include using an Activity Net [74]-like system or active learning, similar to [75]. Relatedly, conducting future keyword searches in languages other than English should help continue to diversify the dataset.

Evaluating the performance of explainability quantitatively requires a new dataset with human-annotated explanations or preferences, along with a benchmark.

Additionally, a crucial component of ASD screening is identifying the absence of certain social behaviors. This underlines the necessity of tools adept at discerning prolonged behavioral absences rather than immediate behavioral presences. To address this in the future, we hope to analyze longer, high-quality videos.

To build a system to monitor the long-term behavioral observation of a child awaiting diagnosis, we may need to go beyond behavior recognition and incorporate localization, tracking, and even anomaly detection. One straightforward idea is to create a two-stage system. The first stage would be a general system focused on detecting/tracking children (both autistic and typical) from other people using a general model. In the second stage, the detected/tracked children's clips could then be further analyzed to identify any autism-related behaviors using our behavior recognition model.

In our next project, we plan to leverage the many neurotypical videos available on social media platforms by approaching autism behavior detection through the lens of anomaly detection [76] or out-of-distribution problem-solving.

Finally, we observe that simply adding more modalities may not always benefit the recognition performance for all behavior categories and sometimes even damage the performance for certain categories. This highlights the trade-off problem of using more modalities for autism behavior recognition. Future models may need to consider how to balance this issue to better leverage multimodal learning.

## VII. CONCLUSION

In this article, we present the AV-ASD dataset, a unique and comprehensive collection featuring social behavioral categories and repetitive behaviors. Our thorough experiments reveal that the integration of visual, audio, and speech data markedly improves autism behavior recognition, thereby facilitating the creation of more effective diagnostic tools. Our LLaVA-ASD model, which combines audio captioning and speech transcription with instruction tuning, excels in utilizing multimodal information for enhanced autism behavior recognition. Additionally, our *post-hoc to ad-hoc* framework represents a pioneering attempt to tackle the challenge of explainability in autism behavior recognition.

## ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to Dr. Carolyn Garver of the Autism Treatment Center for her invaluable suggestions during the project, and to Cristina Rangel Uribe from the Social Communication Lab for her meticulous verification of all annotations to ensure their accuracy.

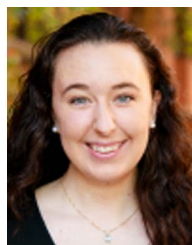
## REFERENCES

- [1] M. B. First, *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*, Washington, DC, USA: American Psychiatric Association, 2022.
- [2] K. Bottema-Beutel, S. K. Kapp, J. N. Lester, N. J. Sasson, and B. N. Hand, "Avoiding ableist language: Suggestions for autism researchers," *Autism Adulthood*, vol. 3, no. 1, pp. 18–29, 2021.
- [3] E. A. Fuller and A. P. Kaiser, "The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis," *J. Autism Devop. Disord.*, vol. 50, pp. 1683–1700, 2020.
- [4] A. S. Nahmias, M. Pellecchia, A. C. Stahmer, and D. S. Mandell, "Effectiveness of community-based early intervention for children with autism spectrum disorder: A meta-analysis," *J. Child Psychol. Psychiatry*, vol. 60, no. 11, pp. 1200–1209, 2019.
- [5] M. van't Hof et al., "Age at autism spectrum disorder diagnosis: A systematic review and meta-analysis from 2012 to 2019," *Autism*, vol. 25, no. 4, pp. 862–873, 2021.
- [6] D. L. Robins, D. Fein, and M. Barton, "M-chat-R/F: The modified checklist for autism in toddlers, revised with follow-up," 2009. [Online]. Available: <https://www.mchatscreen.com/>
- [7] K. Supekar et al., "Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism," *Brit. J. Psychiatry*, vol. 220, no. 4, pp. 202–209, 2022.
- [8] F. Cilia et al., "Computer-aided screening of autism spectrum disorder: Eye-tracking study using data visualization and deep learning," *JMIR Hum. Factors*, vol. 8, no. 4, 2021, Art. no. e27706.
- [9] J. T. Megerian et al., "Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder," *NPJ Digit. Med.*, vol. 5, no. 1, 2022, Art. no. 57.
- [10] W. Jones and A. Klin, "Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, pp. 427–431, 2013.
- [11] P. Mundy and J. Bullen, "The bidirectional social-cognitive mechanisms of the social-attention symptoms of autism," *Front. Psychiatry*, vol. 12, 2022, Art. no. 752274.
- [12] OpenAI, "Hello GPT-4O," May 2024. [Online]. Available: <https://openai.com/index/hello-gpt-4o/>
- [13] C. Li et al., "LaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023, vol. 36, pp. 28541–28564. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/5abdcdf8ecdca cba028c6662789194572-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/5abdcdf8ecdca cba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf)
- [14] E. Billing et al., "The DREAM Dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy," *PLoS one*, vol. 15, no. 8, 2020, Art. no. e0236939.
- [15] J. Rehg et al., "Decoding children's social behavior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3414–3421.
- [16] G. Riva et al., "DE-ENIGMA: Multimodal human-robot interaction for teaching and expanding social imagination in autistic children," *Cyberpsychology, Behav. Social Netw.*, vol. 23, no. 11, pp. 806–807, 2020.
- [17] A. Zunino et al., "Video gesture analysis for autism spectrum disorder detection," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 3421–3426.
- [18] P. Pandey, A. Prathosh, M. Kohli, and J. Pritchard, "Guided weak supervision for action recognition with scarce data to assess skills of children with autism," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 463–470.
- [19] M. D. Coco et al., "Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 4, pp. 993–1004, Dec. 2018.
- [20] G. Dawson et al., "Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder," *Sci. Reports*, vol. 8, no. 1, 2018, Art. no. 17008.
- [21] K. B. Martin et al., "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Mol. Autism*, vol. 9, pp. 1–10, 2018.
- [22] J. Li et al., "MMASD: A multimodal dataset for autism intervention analysis," in *Proc. 25th Int. Conf. Multimodal Interaction*, 2023, pp. 397–405.
- [23] S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 755–761.
- [24] F. Negin, B. Ozyer, S. Agahian, S. Kacdioglu, and G. T. Ozyer, "Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders," *Neurocomputing*, vol. 446, pp. 145–155, 2021.
- [25] P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, and M. A. Armin, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, vol. 9, no. 6, 2023.
- [26] G. O. Ribeiro, M. Grellert, and J. T. Carvalho, "Stimming behavior dataset-unifying stereotype behavior dataset in the wild," in *Proc. IEEE 36th Int. Symp. Comput.-Based Med. Syst.*, 2023, pp. 225–230.
- [27] A. Deng et al., "Problem behaviors recognition in videos using language-assisted deep learning model for children with autism," 2022, *arXiv:2211.09310*.
- [28] A. Ali, F. Negin, S. Thümmel, and F. Bremond, "Video-based behavior understanding of children for objective diagnosis of autism," in *Proc. 17th Int. Conf. Comput. Vis. Theory Appl.*, 2022, pp. 475–484.
- [29] S. S. Rajagopalan and R. Goecke, "Detecting self-stimulatory behaviours for autism diagnosis," in *Proc. 2014 IEEE Int. Conf. Image Process.*, 2014, pp. 1470–1474.
- [30] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5492–5501.
- [31] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," 2020, *arXiv:2001.08740*.
- [32] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10457–10467.
- [33] J. Chen and C. M. Ho, "MM-ViT: Multi-modal video transformer for compressed video action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1910–1921.
- [34] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, Feb. 2012.

- [35] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *IEEE Trans. Multimedia*, vol. 23, pp. 1–11, 2020.
- [36] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 247–263.
- [37] C. Xue, X. Zhong, M. Cai, H. Chen, and W. Wang, "Audio-visual event localization by learning spatial and semantic co-attention," *IEEE Trans. Multimedia*, vol. 25, pp. 418–429, 2021.
- [38] J. Zhou, L. Zheng, Y. Kumar, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8436–8444.
- [39] J. Zhou, D. Guo, and M. Wang, "Contrastive positive sample propagation along the audio-visual event line," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7239–7257, Jun. 2023.
- [40] S. Liu et al., "Dense modality interaction network for audio-visual event localization," *IEEE Trans. Multimedia*, vol. 25, pp. 2734–2748, 2022.
- [41] C. Huang, Y. Tian, A. Kumar, and C. Xu, "Egocentric audio-visual object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22910–22921.
- [42] S. Mo and Y. Tian, "Audio-visual grouping network for sound localization from mixtures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10565–10574.
- [43] Y. Jiang, J. Yin, and Y. Dang, "Leveraging the video-level semantic consistency of event for audio-visual event localization," *IEEE Trans. Multimedia*, vol. 26, pp. 4617–4627, 2023.
- [44] Y.-B. Lin, Y.-L. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2299–2309.
- [45] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 436–454.
- [46] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 34722–34733.
- [47] J. Zhou, D. Guo, Y. Zhong, and M. Wang, "Improving audio-visual video parsing with pseudo visual labels," 2023, *arXiv:2303.02344*.
- [48] J. Zhou et al., "Audio-visual segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 386–403.
- [49] K. Li, Z. Yang, L. Chen, Y. Yang, and J. Xiao, "CATR: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 1485–1494.
- [50] J. Zhou et al., "Audio-visual segmentation with semantics," *Int. J. Comput. Vis.*, pp. 1–21, 2024.
- [51] H. Zhao et al., "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 570–586.
- [52] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15490–15500.
- [53] Y. Su, A. Vosoughi, S. Deng, Y. Tian, and C. Xu, "Separating invisible sounds toward universal audiovisual scene-aware sound separation," 2023, *arXiv:2310.11713*.
- [54] G. Li et al., "Learning to answer questions in dynamic audio-visual scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19108–19118.
- [55] Y. Zhu, Y. Wu, Y. Yang, and Y. Yan, "Describing unseen videos via multi-modal cooperative dialog agents," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 153–169.
- [56] H. Alamri et al., "Audio visual scene-aware dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7558–7567.
- [57] Y. Jiang and J. Yin, "Target-aware spatio-temporal reasoning via answering questions in dynamics audio-visual scenarios," in *Proc. Find. Assoc. Comput. Linguistics*, 2023, pp. 9399–9409.
- [58] H. Zhang, X. Li, and L. Bing, "Video-LLaMa: An instruction-tuned audio-visual language model for video understanding," in *Proc. 2023 Conf. Emp. Methods Natural Lang. Process.: Syst. Demonstr.*, 2023, pp. 543–553.
- [59] F. Shu, L. Zhang, H. Jiang, and C. Xie, "Audio-visual LLM for video understanding, 2023," *arXiv:2312.06720*.
- [60] Z. Cheng et al., "VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs," 2024, *arXiv:2406.07476*.
- [61] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [62] R. Girdhar et al., "ImageBind: One embedding space to bind them all," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15180–15190.
- [63] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.
- [64] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023, vol. 36, pp. 34892–34916. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf)
- [65] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26296–26306.
- [66] W. Kim, C. Choi, W. Lee, and W. Rhee, "An image grid can be worth a video: Zero-shot video question answering using a VLM," 2024, *arXiv:2403.18406*.
- [67] E. Labbé, T. Pellegrini, and J. Pinquier, "CoNeTTE: An efficient audio captioning system leveraging multiple datasets with task embedding," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.
- [68] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 27730–27744.
- [69] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023, vol. 36, pp. 49250–49267.
- [70] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=nZvKeeFYf9>
- [71] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 736–740.
- [72] T. Llama, "The llama 3 herd of models," 2024, *arXiv:2407.21783*.
- [73] S. Chen et al., "VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 72842–72866.
- [74] B. G. Fabian Caba Heilbron, V. Escorcia, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.
- [75] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
- [76] J. Gao, X. Jiang, Y. Yang, D. Li, and L. Qiu, "Unsupervised video anomaly detection for stereotypical behaviours in autism," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.



**Shijian Deng** is currently working toward the Ph.D. degree in computer science with The University of Texas at Dallas, Richardson, TX, USA. His research focuses on leveraging multimodal methods to improve the quality of life for individuals with neurological differences.



**Erin E. Kosloski** received the B.A. degree in psycholinguistics applied to autism from University of Maryland, College Park, MD, USA, minors in spanish and psychology, and a certificate in intercultural communication. She is currently working toward the doctoral degree with Dr. Pamela Rollins' Social Communication Lab, The University of Texas at Dallas, Richardson, TX, USA. Since 2015, she has been working with autistic children and young adults and is passionate about supporting and advocating for people on the autism spectrum. Her research interests include the language, social communication, and joint engagement behaviors of young autistic children as well as the cultural and linguistic responsivity of early interventions for hispanic autistic children and their families.



**Siddhi Patel** is currently working toward the Ph.D. degree with the Social Communication Lab, University of Texas at Dallas, Richardson, TX, USA. She is also a Speech-Language Pathologist. She is interested in exploring the role of early social skills on the developmental outcomes of young autistic children. Her research focuses on social attention and the role of pathways in improving social attention in young autistic children.



**Matthew Wang** is a 12th grade Student with Klein Cain High School, Houston, TX, USA. He is also a Research Intern with the University of Texas at Dallas, Richardson, TX, USA. For undergraduate studies, he plans to go into the field of computer science, possibly with a concentration in artificial intelligence/machine learning. His research interests include the intersection between computer science and humans.



**Zeke A. Barnett** is currently working toward the bachelor's degree in computer science with a concentration in computer graphics with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include multimodal machine learning and 3D graphics engines.



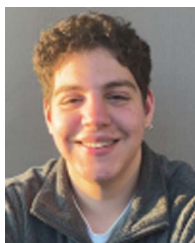
**Harsh Singh** received the bachelor's degree in electronics and communication engineering from IIIT Naya Raipur, India. He is currently working toward the master's degree in computer vision with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, embodied AI, and robotic manipulation.



**Yiyang Nan** received the B.S. degree in statistics and financial mathematics from the University of Michigan, Ann Arbor, MI, USA, in 2021, and the M.S. degree in computer science from Brown University, Providence, RI, USA, in 2024. His research focuses on developing and analyzing capable machine learning models through data-centric approaches.



**Pamela R. Rollins** received the bachelor's degree (*cum laude*) from Boston University, Boston, MA, USA, in 1981, the Masters of Science degree from the University of Wisconsin-Madison, Madison, WI, USA, in 1983, and the Doctor of Education degree from Harvard University, Cambridge, MA, USA, in 1994. She is currently a Professor/Faculty Member with the Department of Speech, Language, and Hearing Sciences, The School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX, USA. Her research has emphasized using behavioral paradigms to understand the dynamics of infant/child social interactions and social experiences as predictors of social attention, social communication, and language development. She extends this work to autistic children, charting developmental trajectories, and conducting translational studies. Her current research focuses on developing culturally and linguistically responsive social communication interventions for low-income Hispanic autistic children, as well as AI applications for assessing social attention.



**Alexander Kaplan** received the bachelor's degree in computer science from the University of Texas at Dallas, Richardson, TX, USA, in 2024. Mr. Kaplan is currently a Software Engineer. Alexander's research interests include deep learning, digital signal processing, and computer networking/streaming. Alexander aims to continue working on technology that helps others and makes a positive impact.



**Yapeng Tian** (Member, IEEE) received the B.E. degree in electronic engineering from Xidian University, Xi'an, China, in 2013, the M.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2017, and the Ph.D. degree in computer science from the University of Rochester, Rochester, NY, USA, in 2022. He is currently an Assistant Professor with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. His research interests include solving core computer vision, computer audition, and machine learning problems and applying the developed learning approaches to broad AI applications, such as multimodal perception, computational photography, AR/VR, accessibility, and healthcare. His work has been recognized with awards, including the AAAI New Faculty Highlights, Cisco Faculty Research Award, and Amazon Research Award.

**Sisira Aarukapalli** photograph and biography not available at the time of publication.



**William T. Doan** is currently working toward the Bachelors of Science degree in computer science on the Texas Top 10% Scholarship. He will graduate early from his program after 2.5 years, in the Fall of 2025. In 2024, as a freshman, he received a grant from the Undergraduate Research Apprenticeship Program. In 2024, he presented his work and was awarded an Erik Jonsson School of Engineering and Computer Science Poster Award for it. His research interests include algorithmic design and analysis, computational complexity, wireless sensor and

ad hoc networks, bioinformatics, string algorithms, and social networks.