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

Shijian Deng, *Member, IEEE,* Erin E. Kosloski, Siddhi Patel, Zeke A. Barnett, Yiyang Nan,
Alexander Kaplan, Sisira Aarukapalli, William T. Doan, Matthew Wang, Harsh Singh,
Pamela R. Rollins, 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, visual, and speech 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. We will release our dataset, code, and pre-trained models.**

***Index Terms*—Audio, Video, Speech, Autism, Multimodal Large Language Model, Explainability, Dataset, Benchmark.**

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[1] children in developing social communication, language, and adaptive functioning [3], [4], yet the worldwide average age at which ASD is diagnosed is 60.48 months [5]. 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.

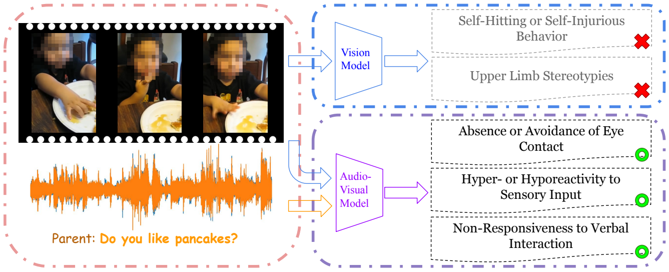

Vision Model
Self-Hitting or Self-Injurious Behavior
Upper Limb Stereotypies
Audio-Visual Model
Absence or Avoidance of Eye Contact
Hyper- or Hypoactivity to Sensory Input
Non-Responsiveness to Verbal Interaction
Parent: Do you like pancakes?

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.

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 RRB 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 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 compre-

S. Deng, Z. Barnett, A. Kaplan, S. Aarukapalli, W Doan, M. Wang, and Y. Tian are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, 75080 USA.

E. Kosloski and P. Rollins are with the School of Behavioral and Brain Sciences, The University of Texas at Dallas, Dallas, TX, 75235 USA .

Y. Nan is with the Department of Computer Science, Brown University, Providence, RI, 02912 USA .

H. Singh is with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE .

[1] 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].

hensive 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-4V [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 that contains 928 clips covering 10 categories annotated from 569 videos, which is the current largest dataset for autism screening in unconstrained videos. The dataset covers 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 by 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 this progress, these datasets primarily focus on RRBs, neglecting the atypical social interactions necessary for autism diagnosis. Leveraging these self-stimulatory behavior and RRB 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.

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	**928**	**10**	✓	✓

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

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], audio-visual video parsing [42], [43], audio-visual segmentation [44], audio-visual source separation [45], [46], [47], and audio-visual question answering or dialogue [48], [49], [50]. 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 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 Tab. 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 ESBD, 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.

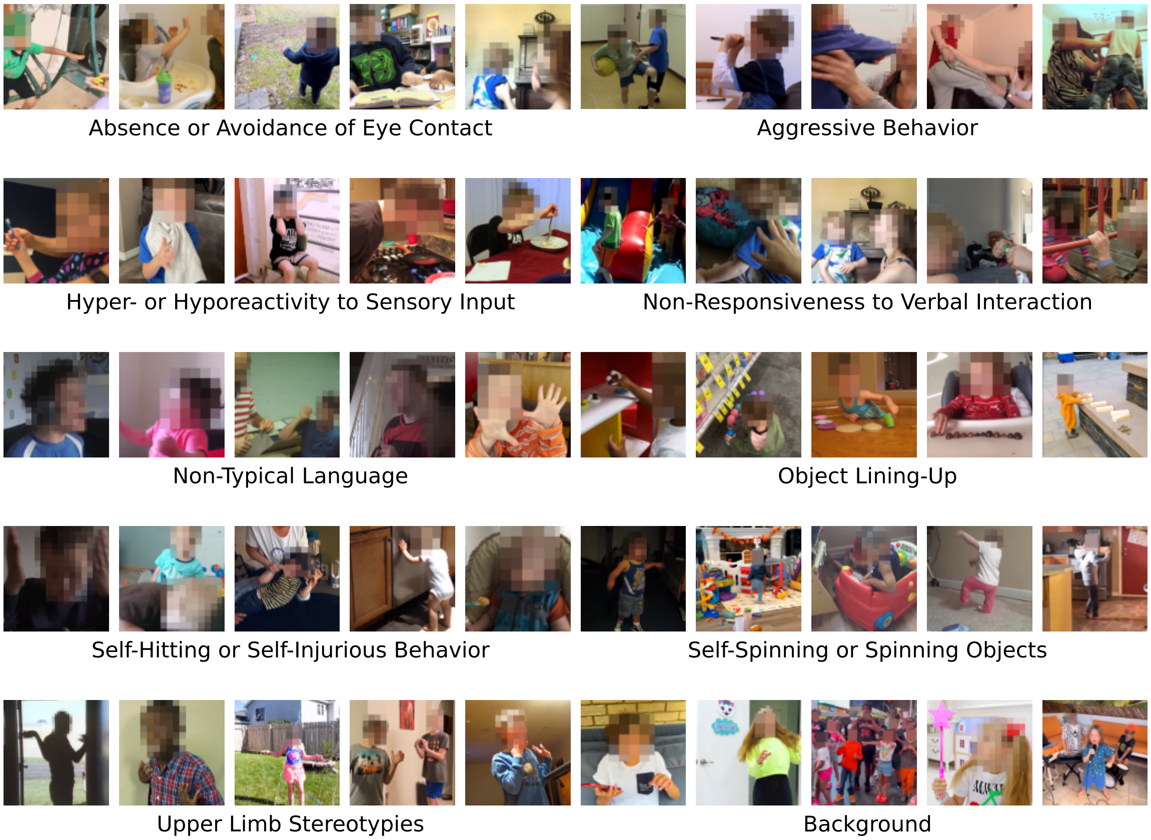

Absence or Avoidance of Eye Contact
Aggressive Behavior
Hyper- or Hyporeactivity to Sensory Input
Non-Responsiveness to Verbal Interaction
Non-Typical Language
Object Lining-Up
Self-Hitting or Self-Injurious Behavior
Self-Spinning or Spinning Objects
Upper Limb Stereotypies
Background

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

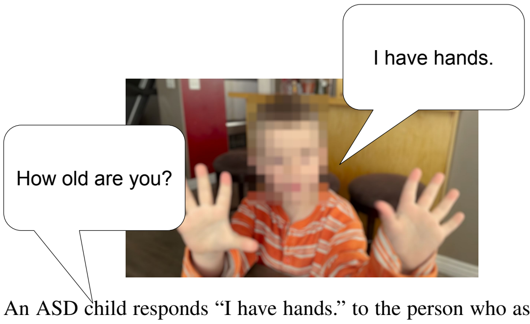


Fig. 3. An ASD child responds "I have hands." to the person who asks "How old are you?"

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

B. Data Collection, Annotation, and Statistics

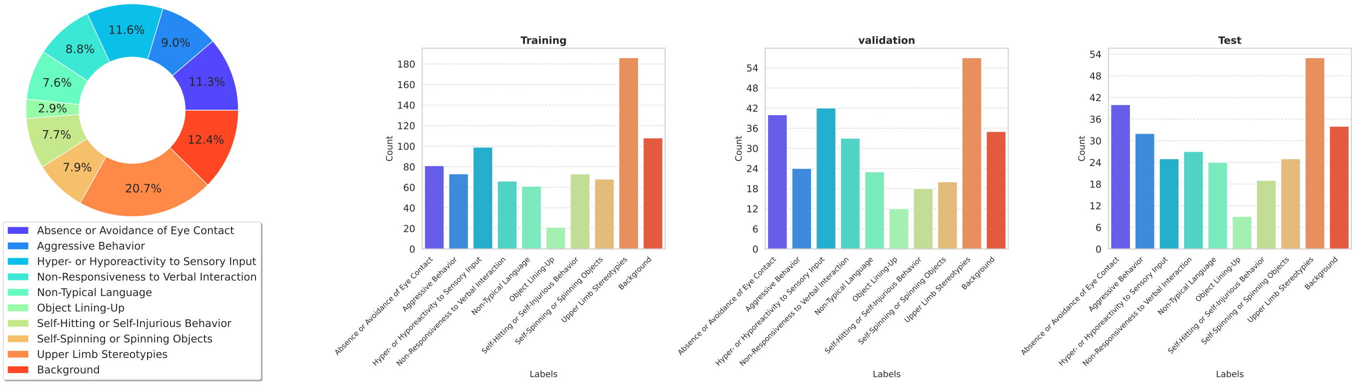
To create our behavioral categories for the social and RRB domains of autism, we identified social challenges from the social behavior classifications of the DSM-V-TR [1] and M-CHAT-R/F [6] screening tool. RRB behaviors were adapted from SSBD [23] and ESBD [24]. Notably, we were limited to behaviors 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 and excluded irrelevant content such as lectures and cartoons. This resulted in 928 distinct video clips extracted from 569 online videos. A team of six volunteer students meticulously annotated each clip, followed by verification by a Speech 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, 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 25.88s, median duration of 10.00s, maximum duration of 887.01s, and minimum duration of 1.00s. Each clip has an average of 1.54 behavior categories.

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.



(a) Category-wise distribution.

(b) Distribution of video in different splits.

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

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 [51] for image features, ImageBind [52] for processing images, videos, and audio, and Whisper [53] 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 $V = \{F_1, F_2, \ldots, F_n\}$, nine frames are uniformly selected and arranged into a 3×3 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 second. 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 [53], 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 $T^m_{t \to t+9} = T(F_{t \to t+9})$. 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^T_m. 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:

$$f = \frac{1}{M} \sum_{m=1}^{M} x_m, \tag{1}$$

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:

$$f = \max_{m=1}^{M} x_m. \tag{2}$$

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

$$f = [x_1, x_2, \ldots, x_M]. \tag{3}$$

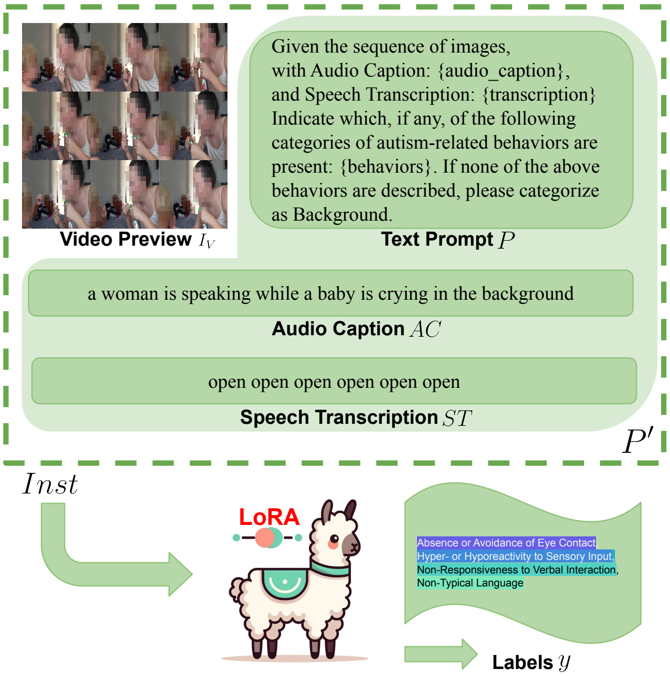

Given the sequence of images,
with Audio Caption: {audio_caption},
and Speech Transcription: {transcription}
Indicate which, if any, of the following
categories of autism-related behaviors are
present: {behaviors}. If none of the above
behaviors are described, please categorize
as Background.
Video Preview I_V
Text Prompt P
a woman is speaking while a baby is crying in the background
Audio Caption AC
open open open open open open
Speech Transcription ST
P'
$Inst$
LoRA
Absence or Avoidance of Eye Contact,
Hyper- or Hyporeactivity to Sensory Input,
Non-Responsiveness to Verbal Interaction,
Non-Typical Language
Labels y

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.

- **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:

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

  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-4V, a state-of-the-art proprietary model developed by OpenAI [12], and LLaVA, an open-source alternative excelling in similar tasks [54], [55] 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. 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 this limitation in our autism behavior recognition task, we propose a Multimodal Representational Text Fusion with two key strategies: (1) Audio Captioning [56]: It transformed audio segments into textual descriptions, enriching the input with semantic information extracted from the audio content; (2) Speech Recognition [53]: This approach transcribed spoken segments into text, providing the model with direct linguistic cues from the audio modality. 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 [57], [58], [59], [60] 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-4V 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 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) [61] 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.

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 [52] 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 [53] 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.

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-4V zero-shot (vision)	16.49
GPT-4V zero-shot (vision, audio)	19.52
GPT-4V zero-shot (vision, speech)	33.88
GPT-4V zero-shot (vision, audio, speech)	28.93
LLaVA zero-shot (vision)	13.26
LLaVA zero-shot (vision, audio)	4.60
LLaVA zero-shot (vision, speech)	15.61
LLaVA zero-shot (vision, audio, speech)	11.61
ImageBind + Whisper (linear probe)	37.36
ImageBind + Whisper (MLP)	39.97
ImageBind + Whisper (temporal)	53.17
Ours (LLaVA-ASD)	**59.77**

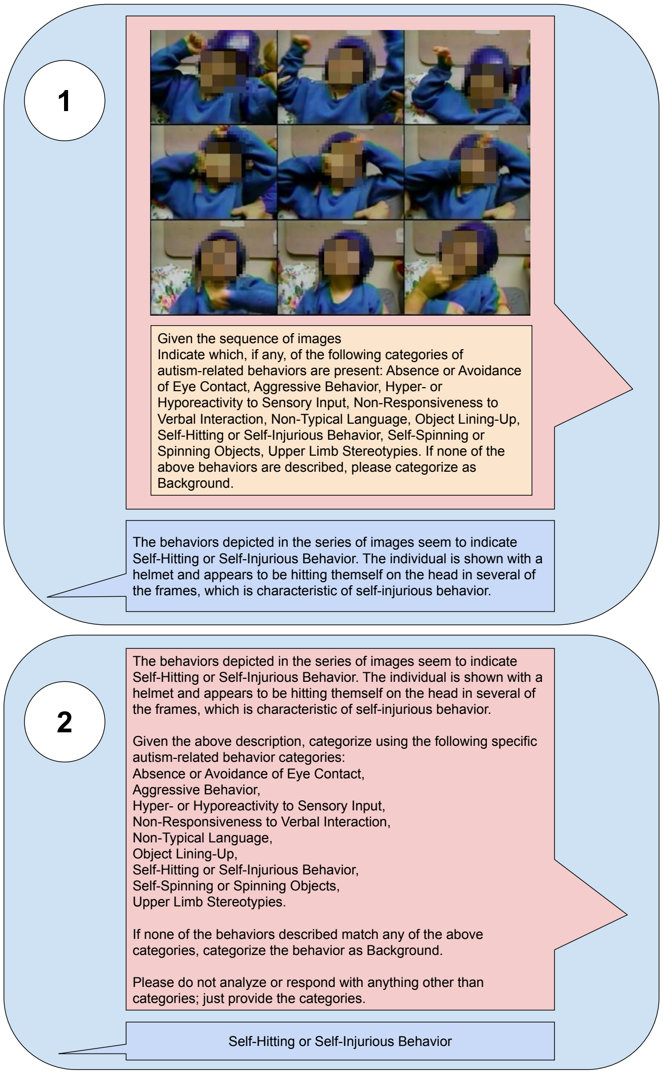
TABLE II

AUTISM BEHAVIOR RECOGNITION RESULTS OF DIFFERENT BASELINES ON AV-ASD TEST SET. THE **BEST** AND SECOND BEST RESULTS ARE HIGHLIGHTED.

We used the CoNeTTE [56], pretrained on Clotho [62], to generate audio captions (AC) and employed Whisper large-v3 [53] 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 1024. Cross-entropy loss with class weights was employed to tackle imbalance issues.



Given the sequence of images
Indicate which, if any, of the following categories of autism-related behaviors are present: Absence or Avoidance of Eye Contact, Aggressive Behavior, Hyper- or Hyporeactivity to Sensory Input, Non-Responsiveness to Verbal Interaction, Non-Typical Language, Object Lining-Up, Self-Hitting or Self-Injurious Behavior, Self-Spinning or Spinning Objects, Upper Limb Stereotypies. If none of the above behaviors are described, please categorize as Background.

The behaviors depicted in the series of images seem to indicate Self-Hitting or Self-Injurious Behavior. The individual is shown with a helmet and appears to be hitting themself on the head in several of the frames, which is characteristic of self-injurious behavior.

The behaviors depicted in the series of images seem to indicate Self-Hitting or Self-Injurious Behavior. The individual is shown with a helmet and appears to be hitting themself on the head in several of the frames, which is characteristic of self-injurious behavior.

Given the above description, categorize using the following specific autism-related behavior categories:
Absence or Avoidance of Eye Contact,
Aggressive Behavior,
Hyper- or Hyporeactivity to Sensory Input,
Non-Responsiveness to Verbal Interaction,
Non-Typical Language,
Object Lining-Up,
Self-Hitting or Self-Injurious Behavior,
Self-Spinning or Spinning Objects,
Upper Limb Stereotypies.

If none of the behaviors described match any of the above categories, categorize the behavior as Background.

Please do not analyze or respond with anything other than categories; just provide the categories.

Self-Hitting or Self-Injurious Behavior

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

MLLMs. To test the zero-shot performance of MLLMs, we used the OpenAI API to get the response of GPT-4V with the specific checkpoint of gpt-4-vision-preview[2]. For LLaVA, we used the Hugging Face checkpoint llava-v1.5-13b[3].

The prompt we used to obtain inference results from GPT-4V 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, oftentimes, 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-4V, allowing it to reformat the answer into a simple behavior label format. This enabled us to easily evaluate performance through word matching.

---

[2] https://platform.openai.com/docs/guides/vision

[3] https://huggingface.co/liuhaotian/llava-v1.5-13b

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	52.70	51.76	51.96	50.83	52.56	**53.17**

TABLE III

F1-score (%) on AV-ASD autism-related behaviors recognition results 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). The last three columns represent weighted ratio sets that we use for the weighted fusion. The **best** and second best results are highlighted.

Behavior	**V**	**V+A**	**V+S**	**V+A+S**
Absence or avoidance of eye contact	46.15	47.89	**56.34**	55.38
Aggressive behavior	72.13	63.33	66.67	**75.00**
Hyper- or hyporeactivity to sensory input	**40.68**	31.17	35.14	29.03
Non-responsiveness to verbal interaction	36.36	33.96	**48.28**	40.00
Non-typical language	20.69	29.27	**45.45**	32.43
Object lining-up	75.00	82.35	85.71	**88.89**
Self-hitting or self-injurious behavior	50.00	40.00	**52.63**	43.90
Self-spinning or spinning objects	56.60	57.69	60.38	**65.38**
Upper limb stereotypies	57.45	58.06	66.02	**67.33**
Background	79.45	**81.69**	81.08	81.01
Average	53.45	52.54	**59.77**	57.84

TABLE IV

Autism behavior recognition results with different modalities by LoRA fine-tuned on LLaVA. The **best** and second best results are highlighted.

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 2×10^{-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.

Evaluation Metric. We used the macro-averaged F1 score to evaluate model performance. For a fair comparison, all models were trained on the same training set of the AV-ASD dataset.

B. Results and Analysis

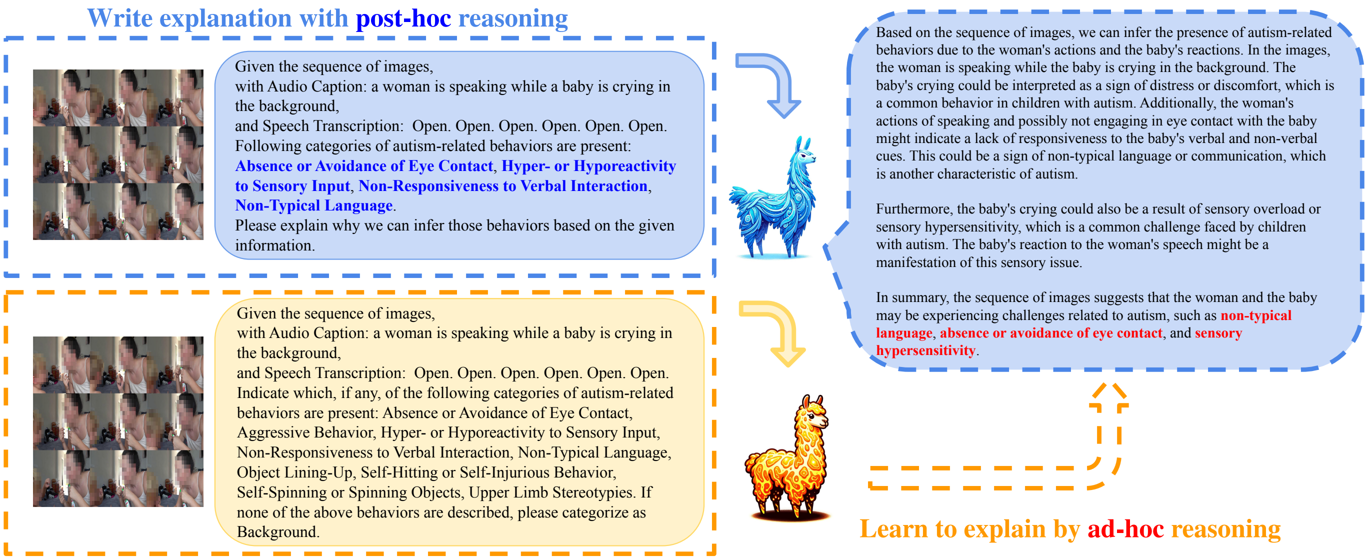
Table II compares the recognition performance of various baselines. Table III provides the ImageBind + Whisper model's performance with different fusion strategies. Table IV provides LLaVA results that show the impact of different modalities.

Vision Matters. The CLIP model achieves an F1 score exceeding 45%, significantly outperforming the dummy baseline, which assigns a positive prediction 1 to all predictions. This indicates that autism behaviors exhibit distinct visual patterns recognizable by vision perception models. Additionally, Table II shows that ImageBind (video, temporal modeling) gains higher accuracy compared to temporal models utilizing audio and speech features alone. These findings underscore the critical role of visual features in autism behavior recognition.

Multimodal integration is helpful. From Table II, we observe that the multimodal model, ImageBind + Whisper, surpasses the performance of unimodal models using solely ImageBind or Whisper features. These results highlight the benefits of integrating cues from audio, visual, and speech modalities in recognizing autism behaviors in videos. Additionally, the findings in Table III and Table IV further substantiate the efficacy of multimodal fusion. Although integrating ambient audio can sometimes be beneficial, we notice it is not a straightforward 'plug and play' solution. For example, when we use audio captions in LLaVA, they do not enhance the overall performance of the task. This suggests that the current audio captions may not be sufficient, and a more effective representation of ambient audio may need to be explored in future work to improve MM-LLM performance in detecting autism-related behaviors.

Zero-shot testing fails. Initially, we tested the zero-shot ability of GPT-4V and LLaVA to identify autistic behavior, but both fell short as in Table II. GPT-4V simply refused to answer most of the requests, and LLaVA predominantly predicted *Background*. The results demonstrate the two MLLMs cannot directly be used to tackle our task.

Instruction tuning with LLaVA works. We conducted further fine-tuning of LLaVA under four settings: vision only, audio-visual, visual-speech, and audio-visual-speech, to evaluate how different modalities contribute to the recognition performance (see Table IV). (1) A noteworthy observation is that although LLaVA's visual encoder is identical to CLIP's, the V-only model's performance (53.45%) significantly surpasses that of the original CLIP equipped with a linear probe (45.03%). This improvement indicates that the LLM component in LLaVA effectively boosts the perception encoder's classifica-


Write explanation with **post-hoc** reasoning
Learn to explain by **ad-hoc** reasoning

Given the sequence of images,
with Audio Caption: a woman is speaking while a baby is crying in the background,
and Speech Transcription: Open. Open. Open. Open. Open. Open.
Following categories of autism-related behaviors are present:
Absence or Avoidance of Eye Contact, **Hyper- or Hyporeactivity to Sensory Input**, **Non-Responsiveness to Verbal Interaction**, **Non-Typical Language**.
Please explain why we can infer those behaviors based on the given information.

Given the sequence of images,
with Audio Caption: a woman is speaking while a baby is crying in the background,
and Speech Transcription: Open. Open. Open. Open. Open. Open.
Indicate which, if any, of the following categories of autism-related behaviors are present: Absence or Avoidance of Eye Contact, Aggressive Behavior, Hyper- or Hyporeactivity to Sensory Input, Non-Responsiveness to Verbal Interaction, Non-Typical Language, Object Lining-Up, Self-Hitting or Self-Injurious Behavior, Self-Spinning or Spinning Objects, Upper Limb Stereotypies. If none of the above behaviors are described, please categorize as Background.

Based on the sequence of images, we can infer the presence of autism-related behaviors due to the woman's actions and the baby's reactions. In the images, the woman is speaking while the baby is crying in the background. The baby's crying could be interpreted as a sign of distress or discomfort, which is a common behavior in children with autism. Additionally, the woman's actions of speaking and possibly not engaging in eye contact with the baby might indicate a lack of responsiveness to the baby's verbal and non-verbal cues. This could be a sign of non-typical language or communication, which is another characteristic of autism.

Furthermore, the baby's crying could also be a result of sensory overload or sensory hypersensitivity, which is a common challenge faced by children with autism. The baby's reaction to the woman's speech might be a manifestation of this sensory issue.

In summary, the sequence of images suggests that the woman and the baby may be experiencing challenges related to autism, such as **non-typical language**, **absence or avoidance of eye contact**, and **sensory hypersensitivity**.

Fig. 7. Explainability Framework of *post-hoc to ad-hoc*. The model initially generates a pseudo-explanation (*Post-Hoc*) based on the provided ground truth. Subsequently, it uses this pseudo-explanation as guidance to learn how to explain the decision-making process in identifying behaviors without the need for provided ground truth (*Ad-Hoc*).

tion efficacy. The fine-tuned model outperformed all previous 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, while combining visual and speech data together achieved 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: Explainility

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 8(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 Figure 8(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. 7. It contains two steps. First, with a visual input, audio caption, speech transcription, and prompt instruction P_{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:

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

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:

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

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

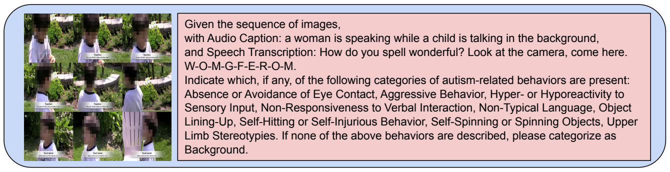
Here, $\hat{P}_{\text{inst}}$ denotes the prompt for the ad-hoc step and $CE(\cdot)$ is cross-entropy loss.

Figure 8(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.

VI. FUTURE DIRECTIONS

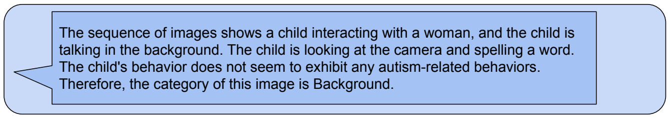
Opportunities for expanding the size of the dataset include using an Activity Net [63]-like system or active learning, similar to [64]. Relatedly, conducting future keyword searches in languages other than English should help continue to diversify the dataset. In our next project, we plan to leverage the many neurotypical videos available on social media platforms by approaching autism behavior detection through the lens of anomaly detection [65] or out-of-distribution problem-solving.

Finally, 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.



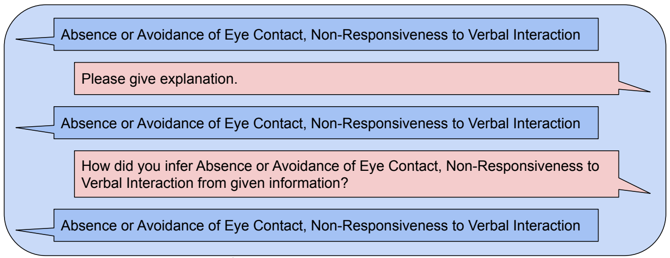
Given the sequence of images,
with Audio Caption: a woman is speaking while a child is talking in the background,
and Speech Transcription: How do you spell wonderful? Look at the camera, come here.
W-O-M-G-F-E-R-O-M.
Indicate which, if any, of the following categories of autism-related behaviors are present: Absence or Avoidance of Eye Contact, Aggressive Behavior, Hyper- or Hyporeactivity to Sensory Input, Non-Responsiveness to Verbal Interaction, Non-Typical Language, Object Lining-Up, Self-Hitting or Self-Injurious Behavior, Self-Spinning or Spinning Objects, Upper Limb Stereotypies. If none of the above behaviors are described, please categorize as Background.

(a) Model Input



The sequence of images shows a child interacting with a woman, and the child is talking in the background. The child is looking at the camera and spelling a word. The child's behavior does not seem to exhibit any autism-related behaviors. Therefore, the category of this image is Background.

(b) Zero-shot Model



Absence or Avoidance of Eye Contact, Non-Responsiveness to Verbal Interaction

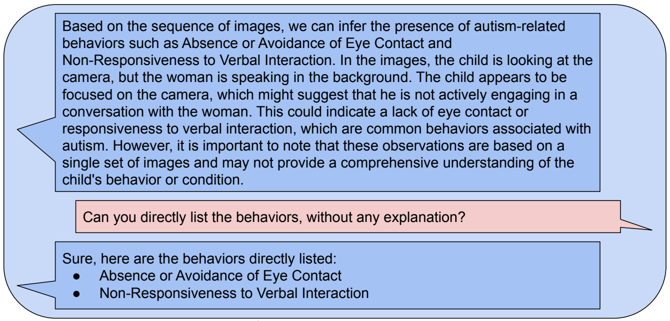
Please give explanation.

Absence or Avoidance of Eye Contact, Non-Responsiveness to Verbal Interaction

How did you infer Absence or Avoidance of Eye Contact, Non-Responsiveness to Verbal Interaction from given information?

Absence or Avoidance of Eye Contact, Non-Responsiveness to Verbal Interaction

(c) Model fine tuned with ground truth behavior labels



Based on the sequence of images, we can infer the presence of autism-related behaviors such as Absence or Avoidance of Eye Contact and Non-Responsiveness to Verbal Interaction. In the images, the child is looking at the camera, but the woman is speaking in the background. The child appears to be focused on the camera, which might suggest that he is not actively engaging in a conversation with the woman. This could indicate a lack of eye contact or responsiveness to verbal interaction, which are common behaviors associated with autism. However, it is important to note that these observations are based on a single set of images and may not provide a comprehensive understanding of the child's behavior or condition.

Can you directly list the behaviors, without any explanation?

Sure, here are the behaviors directly listed:
- Absence or Avoidance of Eye Contact
- Non-Responsiveness to Verbal Interaction

(d) Model fine tuned with pseudo explanation

Fig. 8. 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.

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

We express our 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] APA, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC: 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 in 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," *Journal of autism and developmental disorders*, 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," *Journal of Child Psychology and Psychiatry*, vol. 60, no. 11, pp. 1200–1209, 2019.

[5] M. van't Hof, C. Tisseur, I. van Berckelear-Onnes, A. van Nieuwenhuyzen, A. M. Daniels, M. Deen, H. W. Hoek, and W. A. Ester, "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," Online, 2009, available at: https://www.mchatscreen.com/.

[7] K. Supekar, C. de Los Angeles, S. Ryali, K. Cao, T. Ma, and V. Menon, "Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism," *The British Journal of Psychiatry*, vol. 220, no. 4, pp. 202–209, 2022.

[8] F. Cilia, R. Carette, M. Elbattah, G. Dequen, J.-L. Guérin, J. Bosche, L. Vandromme, B. Le Driant *et al.*, "Computer-aided screening of autism spectrum disorder: Eye-tracking study using data visualization and deep learning," *JMIR human factors*, vol. 8, no. 4, p. e27706, 2021.

[9] J. T. Megerian, S. Dey, R. D. Melmed, D. L. Coury, M. Lerner, C. J. Nicholls, K. Sohl, R. Rouhbakhsh, A. Narasimhan, J. Romain *et al.*, "Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder," *NPJ digital medicine*, vol. 5, no. 1, p. 57, 2022.

[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," *Frontiers in Psychiatry*, vol. 12, p. 752274, 2022.

[12] OpenAI, "Gpt-4 technical report," 2023.

[13] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[14] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. Hernandez Garcia, P. Gómez Esteban *et al.*, "The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy," *PloS one*, vol. 15, no. 8, p. e0236939, 2020.

[15] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3414–3421.

[16] G. Riva, E. Riva *et al.*, "De-enigma: Multimodal human-robot interaction for teaching and expanding social imagination in autistic children," *Cyberpsychology, behavior and social networking*, vol. 23, no. 11, pp. 806–807, 2020.

[17] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 463–470.

[19] M. Del Coco, M. Leo, P. Carcagnì, F. Fama, L. Spadaro, L. Ruta, G. Pioggia, and C. Distante, "Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 993–1004, 2017.

[20] G. Dawson, K. Campbell, J. Hashemi, S. J. Lippmann, V. Smith, K. Carpenter, H. Egger, S. Espinosa, S. Vermeer, J. Baker *et al.*, "Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder," *Scientific reports*, vol. 8, no. 1, p. 17008, 2018.

[21] K. B. Martin, Z. Hammal, G. Ren, J. F. Cohn, J. Cassell, M. Ogihara, J. C. Britton, A. Gutierrez, and D. S. Messinger, "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Molecular autism*, vol. 9, pp. 1–10, 2018.

[22] J. Li, V. Chheang, P. Kullu, E. Brignac, Z. Guo, K. E. Barner, A. Bhat, and R. L. B. Name, "Mmasd: A multimodal dataset for autism intervention analysis," *arXiv preprint arXiv:2306.08243*, 2023.

[23] S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *Proceedings of the IEEE International Conference on Computer Vision 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," *arXiv preprint arXiv:2208.04206*, 2022.

[26] G. O. Ribeiro, M. Grellert, and J. T. Carvalho, "Stimming behavior dataset-unifying stereotype behavior dataset in the wild," in *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2023, pp. 225–230.

[27] A. Deng, T. Yang, C. Chen, Q. Chen, L. Neely, and S. Oyama, "Problem behaviors recognition in videos using language-assisted deep learning model for children with autism," *arXiv preprint arXiv:2211.09310*, 2022.

[28] A. Ali, F. F. Negin, F. F. Bremond, and S. Thümmler, "Video-based behavior understanding of children for objective diagnosis of autism," in *VISAPP 2022-17th International Conference on Computer Vision Theory and Applications*, 2022.

[29] S. S. Rajagopalan and R. Goecke, "Detecting self-stimulatory behaviours for autism diagnosis," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5492–5501.

[31] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[32] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 457–10 467.

[33] J. Chen and C. M. Ho, "Mm-vit: Multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 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 Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, 2011.

[35] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *IEEE Transactions on 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 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 Transactions on Multimedia*, vol. 25, pp. 418–429, 2021.

[38] S. Liu, W. Quan, C. Wang, Y. Liu, B. Liu, and D.-M. Yan, "Dense modality interaction network for audio-visual event localization," *IEEE Transactions on Multimedia*, 2022.

[39] C. Huang, Y. Tian, A. Kumar, and C. Xu, "Egocentric audio-visual object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 910–22 921.

[40] S. Mo and Y. Tian, "Audio-visual grouping network for sound localization from mixtures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 565–10 574.

[41] Y. Jiang, J. Yin, and Y. Dang, "Leveraging the video-level semantic consistency of event for audio-visual event localization," *IEEE Transactions on Multimedia*, 2023.

[42] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 436–454.

[43] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," in *Advances in Neural Information Processing Systems*, 2022.

[44] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, "Audio–visual segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 386–403.

[45] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[46] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 15 490–15 500.

[47] Y. Su, A. Vosoughi, S. Deng, Y. Tian, and C. Xu, "Separating invisible sounds toward universal audiovisual scene-aware sound separation," *arXiv preprint arXiv:2310.11713*, 2023.

[48] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 108–19 118.

[49] Y. Zhu, Y. Wu, Y. Yang, and Y. Yan, "Describing unseen videos via multi-modal cooperative dialog agents," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 153–169.

[50] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson *et al.*, "Audio visual scene-aware dialog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7558–7567.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[52] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.

[53] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[54] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[55] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[56] E. Labbé, T. Pellegrini, and J. Pinquier, "Conette: An efficient audio captioning system leveraging multiple datasets with task embedding," *arXiv preprint arXiv:2309.00454*, 2023.

[57] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[58] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[59] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[60] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[61] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[62] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[63] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[64] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[65] J. Gao, X. Jiang, Y. Yang, D. Li, and L. Qiu, "Unsupervised video anomaly detection for stereotypical behaviours in autism," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.