Audio-Visual Autism Behavior Recognition With Multimodal Large Language Models

Shijian Deng[1] Erin E. Kosloski[1] Siddhi Patel[1] Zeke A. Barnett[1] Yiyang Nan[2] Alexander Kaplan[1]
Sisira Aarukapalli[1] Matthew Wang[1] Harsh Singh[3] Matthew Wang[1] Pamela R. Rollins[1] Yapeng Tian[1]
[1] The University of Texas at Dallas [2] Brown University [3] MBZUAI
{shijian.deng, erin.kosloski, siddhi.patel, zeke.barnett, sasha.kaplan,
sisira.aarukapalli, william.doan, matthew.wang, rollins, yapeng.tian}@utdallas.edu
yiyang_nan@brown.edu harsh.singh@mbzuai.ac.ae

1. 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]. 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 restricted and repetitive behaviors (RRBs) overlooking social interaction challenges integral to the diagnosis of autism. 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.

With the newly collected dataset, we establish a comprehensive benchmark for exploring how to better recognize autism behaviors in videos. We develop several baselines using foundation models like CLIP (image), ImageBind (video/audio), and Whisper (speech). We further investigate the effectiveness of Multimodal Large Language Models (MLLMs), including GPT-4V [11] and LLaVA [6], 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 instruction tuning can compromise the 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.

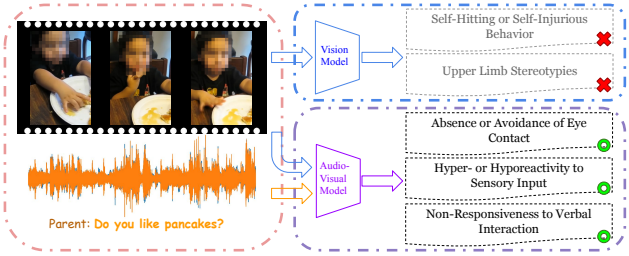

Figure 1. The vision-only model incorrectly identified two behaviors that were not present, whereas the audio-visual model correctly identified the three behaviors. This illustrates how multimodal integration enables more accurate behavior recognition.

2. The AV-ASD Dataset

AV-ASD distinguishes itself from preceding datasets in several significant ways, as delineated in Tab. 1. 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.

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 [16] screening tool. RRB behaviors were adapted from SSBD [14] and ESBD [10]. 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 students

Dataset	Clips	Categories	Multi-Label	Social Behaviors
SSBD [14]	75	3	✗	✗
ESBD [10]	141	4	✗	✗
Wei *et al.* [17]	61	3	✗	✗
ASBD [15]	165	4	✗	✗
AV-ASD	**928**	**10**	✓	✓

Table 1. Comparison of the AV-ASD dataset with other autism-related behavior datasets used in ASD screening research.

meticulously annotated each clip, followed by verification by a Speech Pathologist (SLP) with 15 years of experience working with autistic children. This research utilized data in a manner consistent with previous studies [10, 14], and our university's IRB approved our study.

3. Autism Behavior Recognition

We leverage recent foundation models to extract features for autism behavior recognition, utilizing CLIP [12] for image features, ImageBind [3] for processing images, videos, and audio, and Whisper [13] for speech analysis. These extracted features are then inputted into MLPs for prediction.

To use CLIP and ImageBind as image encoders, a video is transformed into a single composite image. Specifically, given a sequence of frames, nine frames are uniformly selected and arranged into a 3×3 grid to form a composite image denoted as I_V. These extracted features are subsequently utilized in MLP models designed to classify various autism behaviors. Additionally, this composite image format serves as instrumental visual input for MLLMs.

3.1. 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 [11], and LLaVA, an open-source alternative excelling in similar tasks [7, 8] 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. 2). 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 $\hat{y} = \text{MLLM}(I_V, P)$, 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 [5]: It transformed audio segments into textual descriptions, enriching the input with semantic information extracted from the audio content; (2) Speech Recognition [13]: 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 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.

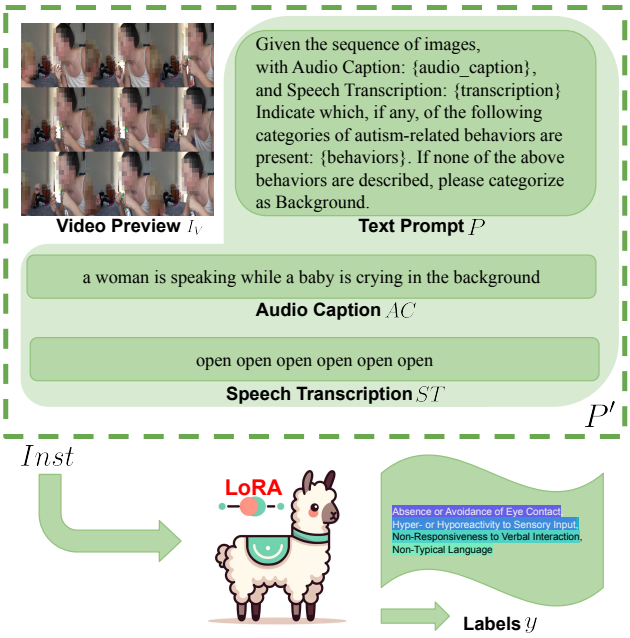

Figure 2. LLaVA-ASD: Instruction Tuning for LLaVA. Given a video preview I_V and an enhanced text prompt P', which is a text prompt P augmented with an audio caption AC and speech transcription ST. These elements are combined to form the model's instruction input $Inst$. The output consists of multiple autism behavior labels presented in text format as y. We employed LoRA for efficient fine-tuning.

3.2. MLLMs with Instruction Tuning

Beyond zero-shot testing, we leverage the power of instruction tuning [2, 9, 18] to enhance MLLMs' performance on our specialized dataset 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 instruction tuning. For training, we construct an instruction tuning pair denoted as $\{Inst, y\}$. Here,

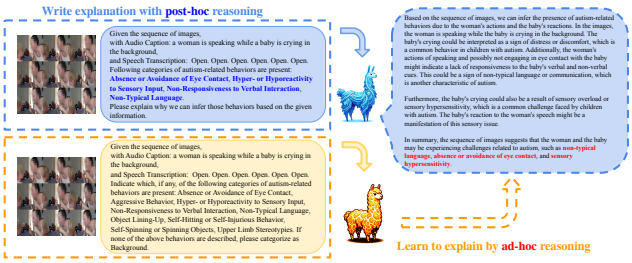

Figure 3. Explainability Framework of *post-hoc to ad-hoc*. The model initially generates a pseudo-explanation (*Post-Hoc*) based on the provided ground truth. Subsequently, it uses this pseudo-explanation as guidance to learn how to explain the decision-making process in identifying behaviors without the need for provided ground truth (*Ad-Hoc*).

$Inst = [I_V, P']$ 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 labels for the behavior categories in the video clip. In implementation, we employed the low-rank adaptation (LoRA) [4] for efficient training. We denote the trained model as LLaVA-ASD.

4. Experiments

4.1. Results and Analysis

MLLMs' 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 2. 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 to evaluate how different modalities contribute to the recognition performance (see Table 3). (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 classification efficacy. The fine-tuned model outperformed all previous baselines listed in Table 2, 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.

4.2. 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. 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. 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. 3. 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:

$$\text{R}_{\text{post-hoc}} = \text{MLLM}(I_V, AC, ST, P_{\text{inst}}, L_{gt}) \quad (1)$$

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:

$$\text{R}_{\text{ad-hoc}} = \text{MLLM}(I_V, AC, ST, \hat{P}_{\text{inst}}; \theta) \quad (2)$$

$$\hat{\theta} = \arg\min_{\theta} CE(\text{R}_{\text{ad-hoc}}, \text{R}_{\text{post-hoc}}) \quad (3)$$

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

Fine-tuning MLLMs with synthetic post-hoc reasoning data effectively prevents the model from reducing to a trivial behavior classifier. Additionally, it improves its explainability for the task of recognizing autism behaviors.

5. Conclusion

In this paper, 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 and speech data markedly improves autism behavior recognition, thereby

Method	Dummy Baseline	CLIP	ImageBind (image)	ImageBind (video)	ImageBind (audio)	Whisper	GPT-4V	LLaVA	Ours (LLaVA-ASD)
F1-score (%)	26.83	45.72	39.19	44.06	28.47	36.48	33.88	15.61	**59.77**

Table 2. Autism behavior recognition results of different baselines on AV-ASD test set. Top-2 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 3. Autism behavior recognition results with different modalities by LoRA fine-tuned on LLaVA.

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.

Acknowledgments. This work was supported in part by UTD SPIRe award. The article solely reflects the opinions and conclusions of its authors but not the funding agent.

References

[1] APA. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, Washington, DC, 5th edition, 2022. 1

[2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[3] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[5] Etienne Labbé, Thomas Pellegrini, and Julien Pinquier. Conette: An efficient audio captioning system leveraging multiple datasets with task embedding. *arXiv preprint arXiv:2309.00454*, 2023. 2

[6] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 1

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[10] Farhood Negin, Baris Ozyer, Saeid Agahian, Sibel Kacdioglu, and Gulsah Tumuklu Ozyer. Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. *Neurocomputing*, 446:145–155, 2021. 1, 2

[11] OpenAI. Gpt-4 technical report, 2023. 1, 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 2

[14] Shyam Rajagopalan, Abhinav Dhall, and Roland Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 755–761, 2013. 1, 2

[15] Guilherme Ocker Ribeiro, Mateus Grellert, and Jonata Tyska Carvalho. Stimming behavior dataset-unifying stereotype behavior dataset in the wild. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 225–230. IEEE, 2023. 2

[16] Diana L. Robins, Deborah Fein, and Marianne Barton. M-chat-r/f: The modified checklist for autism in toddlers, revised with follow-up. Online, 2009. Available at: https://www.mchatscreen.com/. 1

[17] Pengbo Wei, David Ahmedt-Aristizabal, Harshala Gammulle, Simon Denman, and Mohammad Ali Armin. Vision-based activity recognition in children with autism-related behaviors. *arXiv preprint arXiv:2208.04206*, 2022. 2

[18] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2