A Multimodal Self-supervised AI Framework for Monitoring Challenge Behavior Risks in Children with ASD

Zhenhao Zhao[1], Eunsun Chung[2], Kyoug-Mee Chung[2], Michelle Crawford[3], Chung Hyuk Park[1]

[1] *Department of Biomedical Engineering*, *The George Washington University* Washington, DC, USA,
E-mails: zzhao98@gwu.edu, chpark@gwu.edu

[2] *Department of Psychology*, *Yonsei University* Seoul, Korea, E-mails: eun930320@gmail.com, kmchung@yonsei.ac.kr

[3]*Neurobehavioral Unit*, *Kennedy Krieger Institute*, Baltimore, MD, USA, E-mail: crawfordm@kennedykrieger.org

***Abstract*—Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition with increasing prevalence. Its associated challenging behaviors significantly impact social interactions and daily living. Traditional assessments, which rely on clinical evaluations oftentimes requiring long wait times, may easily miss important behavioral episodes and timely training or interventions. This work proposes a novel AI framework that leverages multimodal self-supervised learning to provide at-home monitoring and real-time analysis of interaction styles, to form an embodied AI system that can collaborate with caregivers in the home or clinical settings. Using a two-stage strategy, the framework first extracts clinically meaningful representations of a child's interaction style and then assesses behavioral risks with an interpretable algorithm. On-going efforts involve more rigorous data collection, clinical collaboration, contextual analysis of data outcomes and clinical validation of our algorithm.**

***Index Terms*—Behavior Recognition, Self-supervised learning, Multimodal learning**

I. Introduction

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition characterized by a range of behavioral and cognitive challenges [1]. Over recent decades, the prevalence of ASD among children in the United States has increased dramatically, with current estimates indicating that approximately 1 in 36 children is diagnosed with the disorder [2]–[5]. Among the myriad clinical features associated with ASD, challenging behaviors (CBs)—including self-injurious actions, aggression, and disruptive conduct—stand out due to their profound impact on social interactions and the potential for severe health consequences for both the affected individuals and those around them [6].

Despite their clinical significance, the continuous monitoring of CBs in naturalistic, everyday environments remains a formidable challenge. Traditional assessment approaches largely depend on periodic, in-office clinical evaluations, which not only impose substantial logistical and financial burdens on families but also risk missing transient or sporadic behavioral episodes, potentially leading to diagnostic discrepancies [7]. In light of these limitations, there is a pressing need for an automated, unobtrusive system capable of analyzing home video recordings to capture the nuanced interaction styles of children with ASD, thereby enabling a more accurate and continuous assessment of their risk for CBs.

In this work, we plan to address this challenges by developing a novel artificial intelligence framework. Recent advances in multimodal understanding have demonstrated the potential of AI to integrate and analyze both video and audio data effectively [8]–[10]. However, applying such techniques to the domain of ASD behavior analysis presents two major obstacles. First, home video recordings are inherently noisy: they often suffer from visual disturbances such as camera shake and suboptimal viewing angles, as well as audio interference from environmental sounds and background activities. Although these issues can be partially mitigated through extensive pretraining on large-scale, high-quality annotated datasets, the second challenge—data scarcity—remains significant. Due to privacy concerns and the relative rarity of ASD, publicly available datasets are limited in both size and diversity, restricting the performance of conventional supervised learning models [11]. Combined with our established socially assistive robots for individuals with autism [12]–[15], this proposed research will form a collaborative care-giving support with families of individuals with autism and clinicians.

Inspired by the recent success of multimodal self-supervised learning models, which leverage vast amounts of unannotated data to acquire robust prior knowledge [16], [17], we proposed a multimodal self-supervised AI CBs risk assessment framework. Initially, we utilize our previously developed self-supervised behavior recognition model, AV-FOS [18] , to extract clinically meaningful representations of a child's Interaction Style (IS) as delineated in the Family Observation Schedule (FOS). Building upon these classifications, we propose an interpretable risk assessment algorithm that computes a weighted sum of observed ISs to yield a quantitative risk score. Furthermore, we plan to collect additional data and further validate our model's clinical efficacy by examining

This work was supported in part by the U.S. National Science Foundation (NSF) under Grant #1846658, titled "CAREER: Social Intelligence with Contextual Ambidexterity for Long-Term Human-Robot Interaction and Intervention (LT-HRI^2)".

the correlation between its evaluation outcomes and established clinical assessment instruments, such as the Aberrant Behavior Checklist (ABC) [19] and the Behavior Problems Inventory (BPI-01) [20]. This comprehensive framework not only demonstrates promising performance in preliminary evaluations but also paves the way for improved continuous monitoring and timely interventions in real-world clinical settings.

II. Methods

A. Two Stages Learning for Interaction Styles Recognition

1) FOS Dataset: We utilize the FOS dataset, which comprises multimodal (audio and video) recordings of 83 children with autism and their caregivers. The dataset includes 8,108 ten-second clips, each annotated with 23 ISs derived from the Revised Family Observation Schedule (FOS-R-III). Annotations were performed by trained professionals, achieving an inter-rater agreement exceeding 90%.

2) Tokenization: For each ten-second clip, both visual and audio signals are tokenized to facilitate multimodal learning.

Visual Tokenization: Three uniformly sampled frames are averaged, resized to 224×224, normalized, and partitioned into 16×16 patches, resulting in 196 tokens with added positional and modality embeddings.

Audio Tokenization: The raw waveform (processed at 16 kHz) is converted into a 128-dimensional log Mel-filter bank using a 25 ms Hamming window (10 ms frame shift). Spectrograms are standardized to 1024 frames, segmented into 512 non-overlapping 16×16 patches, and embedded with temporal positional cues.

Figure 1 illustrates the overall tokenization process.

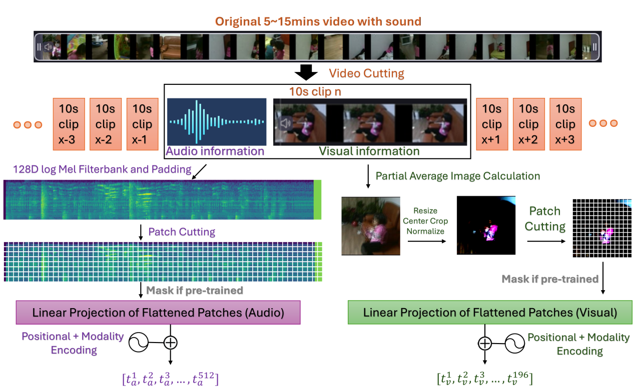

Fig. 1. Overview of the tokenization process.

3) Stage I: Self-supervised Pre-training: We adopt the CAV-MAE framework [21] for self-supervised pretraining. This approach jointly optimizes a cross-modal contrastive loss:

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s^{ii}/\tau)}{\sum_{j=1}^{N}\exp(s^{ij}/\tau)},$$

and a reconstruction loss:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}\|\hat{x}_i - x_i\|_2^2.$$

This dual-objective formulation yields robust cross-modal representations. Figure 2 shows the pretrained structure.

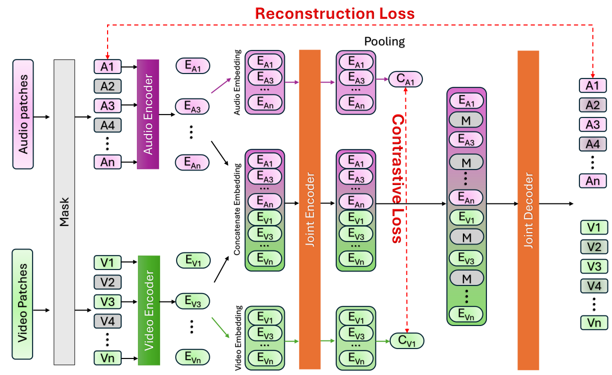

Fig. 2. The CAV-MAE self-supervised learning pretrained structure.

4) Stage II: Supervised Training for IS Classification: After pretraining, we fix the Audio Encoder, Video Encoder, and Joint Encoder as feature extractors. Given the modality-specific token sequences $\mathbf{t}_a$ and $\mathbf{t}_v$, the Joint Encoder E_j fuses the features:

$$\mathbf{Z} = E_j\Big([E_a(\mathbf{t}_a),\ E_v(\mathbf{t}_v)]\Big).$$

We then apply token-level mean pooling:

$$\mathbf{h} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_t,$$

and feed $\mathbf{h}$ into an MLP classifier with a sigmoid activation to generate the final class probability predictions. The overall structure is depicted in Figure 3.

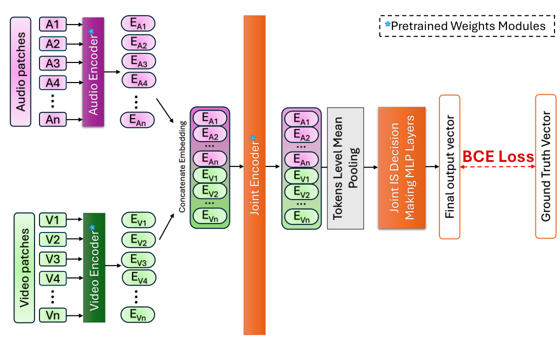

Fig. 3. The AV-FOS self-supervised learning structure.

B. Challenge Behavior Risk Assessment

Upon predicting the interaction styles (IS) from video data, we can subsequently evaluate the risk of challenge behaviors based on these predictions. In the risk evaluation phase, each IS observed in the video is assigned a weight that reflects its empirically and theoretically supported influence on challenge behavior. The overall risk score R is then computed as a weighted sum of the frequencies of the observed IS:

$$R = \sum_{i=1}^{n}\text{Frequency}_i \times \text{Weight}_i,$$

where n represents the total number of interaction styles. A higher risk score is indicative of an increased likelihood of challenge behaviors, whereas a lower score suggests a mitigated risk.

The integration of weighted interaction style features into the risk assessment framework provides a robust methodology for predicting challenge behaviors in autistic children based on observed interaction patterns.

C. Clinical Validation and Contextual Analysis of Data Outcomes

To assess the clinical validity of our AI-based risk evaluation algorithm, we propose a study that investigates the correlation between the algorithm-generated risk scores and established clinical assessment tools. In particular, we focus on the Aberrant Behavior Checklist (ABC) and the Behavior Problems Inventory (BPI-01), both of which are widely recognized for their clinical interpretability in evaluating challenging behaviors in children with ASD. The ABC is a caregiver-completed questionnaire that typically yields a total score ranging up to a maximum value (e.g., 174 for the full-scale score), with higher scores indicating a greater propensity for challenging behaviors. Similarly, the BPI-01, which assesses self-injurious, stereotyped, and aggressive behaviors, produces higher scores when the severity of behavioral problems increases.

For the clinical validation, we plan to collect data from a substantial cohort of ASD children. For each participant, our algorithm will generate a risk score, denoted as R_{AI}, based on the analysis of multimodal interaction data. Concurrently, caregivers will complete both the ABC and the BPI-01 questionnaires, yielding clinical scores that serve as ground truth indicators of behavioral risk. For each subject, we thus obtain a paired score (R_{AI}, Y), where R_{AI} represents the risk score computed from our algorithm, and Y represents the clinical assessment score, which is derived from either R_{ABC} or R_{BPI}.

To rigorously evaluate the relationship between R_{AI} and the clinical assessment scores, we will compute the Pearson correlation coefficient ρ defined as:

$$\rho(R_{\text{AI}}, Y) = \frac{\sum_{i=1}^{N} \left(R_{\text{AI}}^{(i)} - \overline{R_{\text{AI}}}\right)\left(Y^{(i)} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(R_{\text{AI}}^{(i)} - \overline{R_{\text{AI}}}\right)^2} \sqrt{\sum_{i=1}^{N} \left(Y^{(i)} - \overline{Y}\right)^2}},$$

where N is the number of subjects, $\overline{R_{\text{AI}}}$ is the mean of the AI risk scores, and $\overline{Y}$ is the mean of the clinical scores, with Y coming from either R_{ABC} or R_{BPI}. A statistically significant and strong positive correlation (i.e., ρ close to 1) would indicate that higher risk scores predicted by our algorithm are associated with higher clinical ratings of challenging behaviors, thereby supporting the clinical applicability of our approach.

This methodology not only validates the predictive performance of the proposed algorithm in a clinical context but also facilitates a deeper contextual analysis of data outcomes, reinforcing the interpretability and robustness of our risk assessment framework.

III. Initial Results

At the current stage, we have successfully completed the training of the AV-FOS model, which exhibits remarkable performance in recognizing interaction styles (IS). To evaluate its effectiveness, we benchmarked the AV-FOS model against state-of-the-art video understanding architectures, namely the SlowFast Network and the Vision Transformer, as well as against GPT4V with prompt. Table I and Fig. 4 summarize the comparative performance.

TABLE I
AV-FOS Performance on the FOS Dataset.

Model	mAP ↑	Accuracy ↑	Strict Accuracy ↑	AUC ↑	F1 Score ↑	Time Cost →
GPT4V + Prompt V1	0.3181	0.7965	0.1355	0.6624	0.4581	4.3349
GPT4V + Prompt V2	0.2481	0.7668	0.1468	0.5896	0.3330	3.9792
SlowFast	0.6138	0.8287	0.1125	0.8445	0.5437	0.0031
ViT	0.6167	0.8172	0.0889	0.8486	0.5448	0.0011
AV-FOS Model	**0.6879**	**0.8590**	**0.2003**	**0.8868**	**0.5936**	**0.0027**

Our AV-FOS model demonstrates exceptional capabilities, outperforming the competing models in key metrics such as mean Average Precision (mAP), accuracy, Area Under the Curve (AUC), and F1 Score. These results highlight the robustness and predictive accuracy of the AV-FOS approach in capturing and utilizing IS features for effective challenge behavior risk assessment.

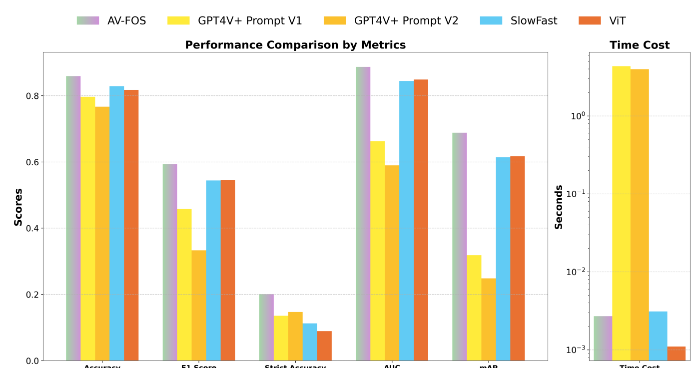

Fig. 4. Performance and time cost comparison among evaluated models.

At the same time, we conducted a preliminary exploration of the impact of IS on the CB risk. Table II presents our analysis of the influence of each Interaction Style on the final CB risk, along with their corresponding weights.

IV. Conclusion and Future Work

In this work, we introduced a novel multimodal self-supervised AI framework for the continuous assessment of challenging behavior risk in children with ASD. Our approach leverages a two-stage strategy, where the first stage employs a self-supervised behavior recognition model (AV-FOS) to extract clinically meaningful representations of a child's interaction style, and the second stage applies an interpretable risk assessment algorithm that computes a weighted sum of observed interaction styles to yield a quantitative risk score. Preliminary evaluations demonstrate promising performance,

TABLE II
SUGGESTED WEIGHTS FOR INTERACTION STYLES IN PREDICTING THE RISK OF CHALLENGE BEHAVIORS AMONG AUTISTIC CHILDREN

IS Code	IS Name	Weight	Rationale
AD	Adhesive Demand	+0.4	Demand-based interactions may induce stress responses and cognitive overload, thereby potentially eliciting challenge behaviors.
AV	Appropriate Verbal Interactions	-0.6	Appropriate verbal interactions provide emotional support and structured guidance, effectively mitigating the risk of emotional dysregulation.
Aff_child	Children Affection	-0.8	Peer affection fosters emotional regulation and a sense of security, thereby reducing the propensity for challenge behaviors.
Aff_parent	Parent Affection	-0.7	Parental affection reinforces secure attachment bonds, thus decreasing negative affect and resistance behaviors.
C+	Positive Contact	-1.0	Positive physical contact, such as hugging or gentle touch, has been shown to significantly lower anxiety levels and enhance emotional stability.
C-	Negative Contact	+1.0	Negative physical contact can result in discomfort and increased stress, thereby heightening the likelihood of challenge behaviors.
CP	Complaint	+0.6	The expression of complaints may indicate underlying distress and communication difficulties, which could elevate behavioral risks.
EA	Engaged Activity of Play	-1.0	Engagement in play activities distracts from negative affect and promotes self-regulation, substantially lowering the risk of challenge behaviors.
Int_child	Children Interrupt	+0.2	Interruptions by children may reflect impulsivity; however, they can also indicate active engagement, hence a modest increase in risk.
Int_parent	Parent Interrupt	+0.3	Parental interruptions may disrupt interactional continuity, slightly increasing risk by limiting opportunities for emotional expression.
MI	Multiple Instructions	+0.4	The provision of multiple instructions may lead to cognitive overload, inducing confusion and subsequent resistance behaviors.
NC	Non-compliance	+0.8	Non-compliant behaviors are directly associated with challenge behaviors, reflecting significant difficulties in emotional regulation.
O	Opposition	+0.9	Oppositional behaviors serve as key indicators of escalating conflict and emotional dysregulation, thus predicting challenge behaviors.
P	Praise	-0.7	Praise positively reinforces adaptive behaviors, enhancing self-confidence and thereby reducing the likelihood of behavioral challenges.
PN	Physical Negative	+1.0	Negative physical interactions, such as hitting or pushing, directly trigger adverse emotional responses, markedly increasing risk.
Q+	Positive Question	-0.4	Positive questioning promotes engagement and reflection, which helps to reduce misunderstandings and emotional tension.
Q-	Negative Question	+0.5	Negative questioning may evoke defensive responses, thereby exacerbating oppositional attitudes.
S+	Positive Social Attention	-1.0	Positive social attention engenders feelings of acceptance and understanding, substantially stabilizing emotional responses.
S-	Negative Social Attention	+0.5	Negative social attention may intensify self-doubt and trigger adverse behavioral reactions.
SI+	Positive Specific Instruction	-0.3	Clear, positively framed instructions facilitate comprehension and task engagement, albeit with a moderate effect.
SI-	Negative Specific Instruction	+0.6	Negative specific instructions can evoke feelings of rejection and undue pressure, leading to behavioral dysregulation.
VI+	Positive Vague Instruction	-0.2	Although vague, positively-toned instructions still offer limited emotional support, resulting in a minor protective effect.
VI-	Negative Vague Instruction	+0.4	Ambiguous and negatively framed instructions may create confusion and cognitive strain, moderately increasing risk.

highlighting the framework's potential for effective at-home monitoring and timely intervention.

Moving forward, our future work will focus on several key directions. First, we plan to augment our dataset with additional high-quality recordings to address the inherent data scarcity and improve model generalizability. Second, we will conduct rigorous clinical validation by correlating our AI-generated risk scores with established clinical assessment instruments, such as the ABC and the BPI-01. This validation will assess the clinical efficacy of our model. Third, we also plan to refine the framework by integrating more comprehensive contextual analysis and exploring potential extensions to accommodate a broader spectrum of behavioral phenotypes, thereby further bridging the gap between AI-driven analysis and clinical practice. Lastly, we will realize this framework through our social robotic platforms to form a closed-loop embodied AI system to provide timely and *in-situ* support in home or clinical settings.

V. ACKNOWLEDGMENTS

This project was partially supported by the NSF Projects #1846658: "CAREER: Social Intelligence with Contextual Ambidexterity for Long-Term Human-Robot Interaction and Intervention (LT-HRI2)" and #2348081: "Collaborative Re-

search: Designing Intelligent Industrial Robots for STEM Inclusion by Leveraging Self-Determination Theory to Foster Autistic Talent in Manufacturing Work."

References

[1] M. M. Hughes, K. A. Shaw, M. DiRienzo, M. S. Durkin, A. Esler, J. Hall-Lande, L. Wiggins, W. Zahorodny, A. Singer, and M. J. Maenner, "The prevalence and characteristics of children with profound autism, 15 sites, united states, 2000-2016," *Public Health Reports®*, vol. 138, no. 6, pp. 971–980, 2023, pMID: 37074176. [Online]. Available: https://doi.org/10.1177/00333549231163551

[2] E. Harris, "Autism Prevalence Has Been on the Rise in the US for Decades—And That's Progress," *JAMA*, vol. 329, no. 20, pp. 1724–1726, 05 2023. [Online]. Available: https://doi.org/10.1001/jama.2023.6078

[3] J. V. Smith, M. Menezes, S. Brunt, J. Pappagianopoulos, E. Sadikova, and M. O. Mazurek, "Understanding autism diagnosis in primary care: Rates of diagnosis from 2004 to 2019 and child age at diagnosis," *Autism*, vol. 0, no. 0, p. 13623613241236112, 0, pMID: 38456360. [Online]. Available: https://doi.org/10.1177/13623613241236112

[4] National Autism Association, "Autism fact sheet," 2024, accessed: 2024-05-22. [Online]. Available: https://nationalautismassociation.org/resources/autism-fact-sheet/

[5] M. J. Maenner, Z. Warren, A. R. Williams, and et al., "Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2020," *MMWR Surveill Summ*, vol. 72, no. No. SS-2, pp. 1–14, 2023.

[6] S. D. Mayes and S. L. Calboun, "Symptoms of autism in young children and correspondence with the dsm," *Infants & Young Children*, vol. 12, no. 2, pp. 11–23, 1999.

[7] T. A. Lavelle, M. C. Weinstein, J. P. Newhouse, K. Munir, K. A. Kuhlthau, and L. A. Prosser, "Economic burden of childhood autism spectrum disorders," *Pediatrics*, vol. 133, no. 3, pp. e520–9, 2014, epub 2014 Feb 10.

[8] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-08836-3

[9] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[10] R. Karim and R. P. Wildes, "Understanding video transformers for segmentation: A survey of application and interpretability," 2023.

[11] P. U. Ravva, B. Kiafar, P. Kullu, J. Li, A. Bhat, and R. L. Barmaki, "Mmasd+: A novel dataset for privacy-preserving behavior analysis of children with autism spectrum disorder," *arXiv preprint arXiv:2408.15077*, 2024.

[12] R. Bevill, C. H. Park, H. J. Kim, J. W. Lee, A. Rennie, M. Jeon, and A. M. Howard, "Interactive robotic framework for multi-sensory therapy for children with autism spectrum disorder," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 421–422.

[13] H. Javed, M. Jeon, A. Howard, and C. H. Park, "Robot-assisted socio-emotional intervention framework for children with autism spectrum disorder," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 131–132.

[14] H. Javed and C. H. Park, "Promoting social engagement with a multi-role dancing robot for in-home autism care," *Frontiers in Robotics and AI*, vol. 9, p. 880691, 2022.

[15] B. Xie and C. H. Park, "An empathetic social robot with modular anxiety interventions for autistic adolescents," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 1148–1155.

[16] Y. Tay, J. Wei, H. W. Chung, V. Q. Tran, D. R. So, S. Shakeri, X. Garcia, H. S. Zheng, J. Rao, A. Chowdhery, D. Zhou, D. Metzler, S. Petrov, N. Houlsby, Q. V. Le, and M. Dehghani, "Transcending scaling laws with 0.1

[17] A. Sengupta, Y. Goel, and T. Chakraborty, "How to upscale neural networks with scaling law? a survey and practical guidelines," 2025. [Online]. Available: https://arxiv.org/abs/2502.12051

[18] Z. Zhao, E. Chung, K.-M. Chung, and C. H. Park, "Av-fos: A transformer-based audio-visual multi-modal interaction style recognition for children with autism based on the family observation schedule (fos-ii)," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–18, 2025.

[19] M. G. Aman, N. N. Singh, A. W. Stewart, and C. Field, "The aberrant behavior checklist: a behavior rating scale for the assessment of treatment effects." *American journal of mental deficiency*, vol. 89, no. 5, pp. 485–491, 1985.

[20] J. Rojahn, J. L. Matson, D. Lott, A. J. Esbensen, and Y. Smalls, "The behavior problems inventory: An instrument for the assessment of self-injury, stereotyped behavior, and aggression/destruction in individuals with developmental disabilities," *Journal of autism and developmental disorders*, vol. 31, pp. 577–588, 2001.

[21] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass, "Contrastive audio-visual masked autoencoder," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=QPtMRyk5rb