Multimodal and Multi-Lingual Deep Neural Network for Interactive Behavior Style Recognition from Uncontrolled Video-logs of Children with Autism Spectrum Disorder

Zhenhao Zhao[1], Eunsun Chung[2], Myungeun Lee[3], Kyong-Mee Chung[2], Chung Hyuk Park[1,4]

***Abstract*— As Autism Spectrum Disorder (ASD) diagnoses rise, revealing underserved populations, the need for efficient public health support is crucial. The Family Observation Schedule-Second Version (FOS-II) is one of the key methods in assessing parent-child interactions in developmental disabilities, yet its manual annotation is time-consuming. This study proposes a multimodal AI model using video input for automated FOS-II annotation. Utilizing advanced deep learning for behavior recognition, this method offers rapid, cost-effective FOS-II scaling. It enhances the capability of social assistive robots to understand human behavior and supports the advancement of digital health research for children with ASD. Of key importance, the visual perception in home settings are most likely based on uncontrolled environments, so it is crucial to develop algorithms that can robustly work with video-log data with uncontrolled quality. Ultimately, it aims to ease the burden on parents and caregivers, streamlining the monitoring and treatment of challenging behaviors in ASD.**

I. INTRODUCTION

The prevalence of Autism Spectrum Disorder (ASD) has been steadily increasing over the past decades, revealing many deficiencies in societal support and meeting family needs globally [1], [2], [3]. Current statistics in the United States indicate that 1 in 36 children is diagnosed with ASD [4]. Children with ASD often show challenging behaviors such as self-injurious behaviors, aggression, disruptive behaviors, and stereotypy, which greatly impact daily social functioning [5].

The Family Observation Schedule-Second Version (FOS-II) [6] is a direct observation tool for assessing parent-child interactions across various contexts. Within the realm of ASD research, FOS-II is often used in clinical and research settings in identifying and evaluating parent-child interactions especially for challenging behaviors, offering valuable insights for providing intervention, and support of children with ASD by examining their social contexts and dynamics [7]. Presently, FOS-II data is encoded manually by trained observers, which is a time-consuming and labor-intensive process. An automated FOS-II encoding algorithm could alleviate the burden on clinicians and researchers, ultimately benefiting numerous children with ASD and their families.

For this, our study proposes an effective multi-modal deep learning model for the automated encoding of FOS-II labels, utilizing videos of children with ASD engaged in specific tasks as training data. The multi-modal data incorporated in this research includes vision and language data, presenting an interesting problem of behavioral understanding with multi-modal data in home settings. In our previous works, we have developed a robotic framework capable of playing interactive social games with autistic children while utilizing mulitmodal perception[8], [9], [10], [11], [12], [13]. However, the behavioral perception was based on general knowledge without any domain-specific clinical knowledge, which limited its ability to comprehensively understand the diverse and sometimes challenging behaviors of children with ASD. We believe that a multimodal deep learning-based approach with clinical-knowledge infused data will enhance the robot's capability to understand human behavior more effectively. And in previous research [14], [15], [16], [17], the behavior recognition of ASD children has predominantly focused on single modalities, either visual or linguistic. Furthermore, the perception of the visual modality has been primarily directed towards identifying individual frames or images in videos, thus fully utilizing spatial information but neglecting valuable temporal information, resulting in less precise perception. The present study not only utilizes

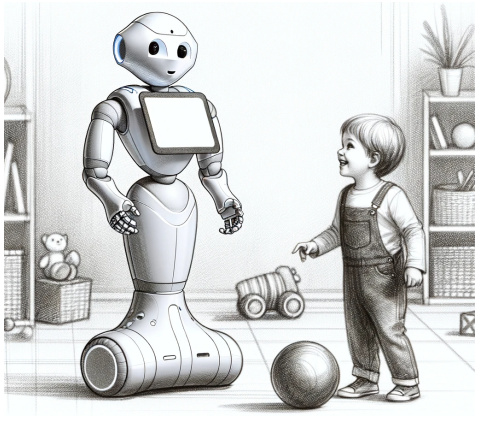

Fig. 1. An illustration showing our vision for social-robot's role in home-based interaction and interventions.

[1]Department of Biomedical Engineering, School of Engineering and Applied Science, The George Washington University zzhao98@gwu.edu; chpark@gwu.edu

[2]Department of Psychology, College of Liberal Arts, Yonsei University eun930320@gmail.com; kmchung@yonsei.ac.kr

[3]Department of Electrical and Computer Engineering, School of Engineering and Applied Science, The George Washington University myungeunlee@gwu.edu

[4]Department of Computer Science, School of Engineering and Applied Science, The George Washington University

the advanced SlowFast architectures[18] to fully exploit both spatial and temporal information in videos but also introduces recognition through the linguistic modality. Additionally, the perception integration through a fully connected layer enables the capture of associations between linguistic and visual modalities, offering more accurate behavior predictions. This aspect has consistently been overlooked in prior ASD behavior recognition research. Moreover, the model trained in this study can accurately identify interaction patterns between children with ASD and their caregivers, parents, or healthcare providers, as opposed to models that solely recognize the behaviors of ASD children themselves [14], [15]. The algorithm proposed in this paper will enhance the ability of social assistive robots to understand human behavior and boost the development of digital health for the children with ASD research.

II. Methods

In this study, we propose a deep learning-based model for recognizing the behavioral and contextual **interaction styles (IS)** between children with ASD and their guardians/healthcare providers. The model, founded on the state-of-the-art artificial intelligence frameworks, can categorize the IS in real-time through videos of children with ASD and their guardians/healthcare providers. The videos are pre-processed and divided into two parts for input into the model:

- The video information, in RGB color domains, is directly provided into the Vision Perception Module
- The audio information, after being processed and converted into text, is fed into the Language Perception Module.

Both modules independently generate vision and language feature vectors. Finally, these two feature vectors are fused through a fully connected neural network for fusion perception, resulting in the prediction of the IS. The entire model structure is illustrated in Figure 2.

A. Dataset Description

Our dataset (FOS-II dataset) includes 216 videos from total 83 ASD children with or without their parents/guardians (caregivers). There are some variants, but mostly composed 3 sets of 10 minutes. The camera was not fixed; instead, a trained researcher held it and recorded to ensure both child and the caregivers were presented in the frame. A wide range of children and parents were included, with severity of problem behaviors ranging from mild to severe. The involving children ages are between 1 to 12 years old. It may cause some bias to our deep learning model based on ages changes, but in the future, we can collect more data with elder children to solve this issue.

The children in the videos performed one of three different tasks: (1) playing with specific types of toys, (2) performing a series of specific instructions, and (3) free playing alone. These videos are encoded every 10 seconds to describe the IS of the children and their parents during they perform task. These ISs are used as labels for training deep learning model. This dataset encompasses 23 types of ISs, including descriptions of caregivers IS (Praise (P) and Affection (AF)), as well as child IS (Non-compliance (NC) and Opposition (O)). Certain IS types are denoted with positive or negative symbols to characterize the emotional tone behind the IS; for example, SA+ signifies positive social attention, whereas SA- denotes negative social attention. For a comprehensive overview of IS types, refer to the table I. In addition, to safeguard the privacy of participants in the dataset, we will not display sample data in the form of images.

IS Code	IS Name
AD	Adersive demand
AV	Appropriate verbal interactions
Aff_child	Children affection
Aff_parent	Parent affection
C+	Positive contact
C-	Negative contact
CP	Complaint
EA	Engaged activity of play
Int_child	Children interrupt
Int_parent	Parent interrupt
MI	Multiple instructions
NC	Non-compliance
O	Opposition
P	Praise
PN	Physical negative
Q+	Positive question
Q-	Negative question
S+	Positive social attention
S-	Negative social attention
SI+	Positive specific instruction
SI-	Negative specific instruction
VI+	Positive vague instruction
VI-	Negative vague instruction

TABLE I
THE EXPLANATION OF EACH IS

The IS human coders are the five trained graduate students from the department of psychology at Yonsei University under the supervision of a clinical psychologist with the license and a Board Certified Behavior Analyst (BCBA) certification. During the FOS coding, Human coders worked in pairs to measure inter-observer reliability. They were trained until they achieved an inter-observer agreement of at least 80%. Subsequently, we consolidated annotations from five annotators to determine the final gold standard.

B. Data Pre-processing

The original video-logs have been coded in every 10-second segment, with each segment having a set of corresponding labels, the IS. One 10-second segment may have multiple ISs instead of one. The 10s segment has been separated into 2 parts: RGB vision information and audio information.

Each 10-second segment of RGB vision data, consisting purely of video, is processed through two different frame sampling frequencies—sparse and dense—to obtain two sets of keyframes. Specifically, we first sample 32 frames from the original video. Then, we duplicate 4 frames from these 32 frames as slow pathway frames, and the remaining 32 frames undergo resolution reduction to serve as fast pathway frames.

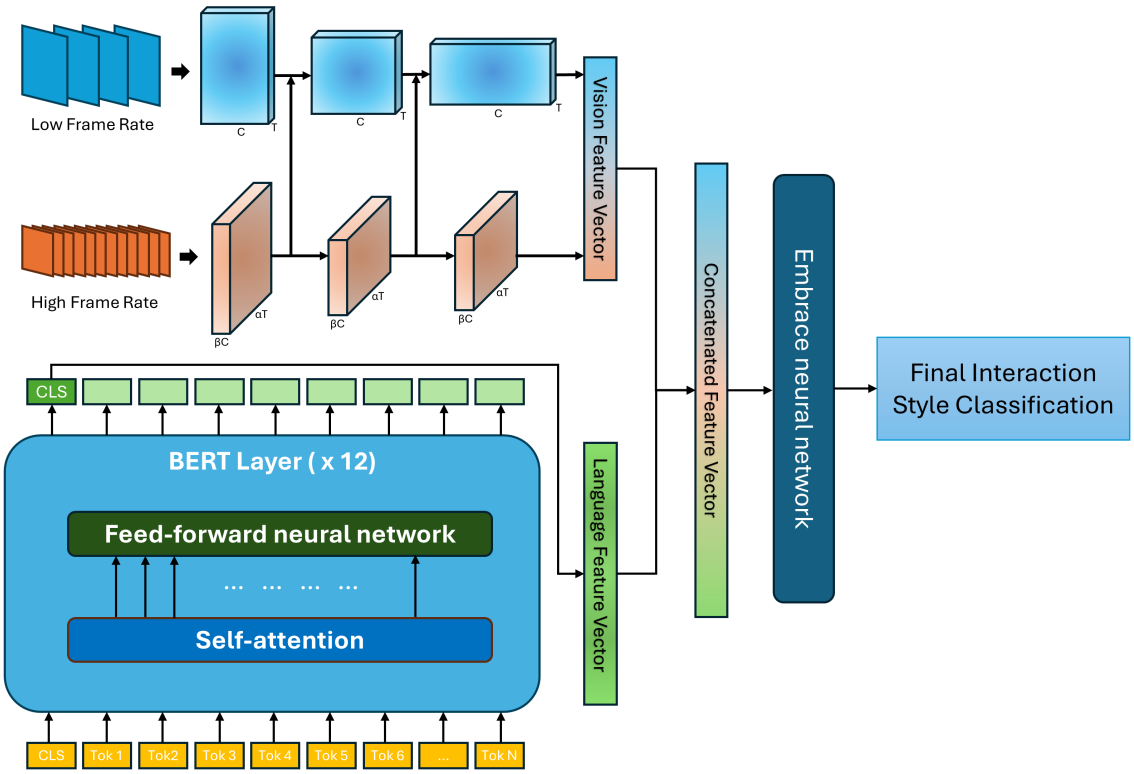

Fig. 2. The multimodal automatic annotation model for the Family Observation Schedule-II (FOS-II)[6] scale proposed in this study is based on the BERT [19] and SlowFast [18] architectures. This model is capable of automatically recognizing the behavioral and contextual **interaction styles (IS)** of children with ASD, thereby alleviating the burden on caretakers

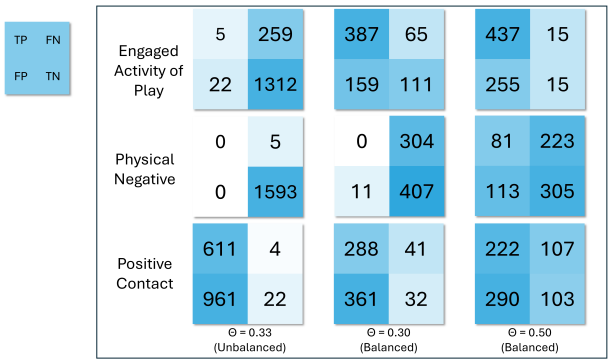

Fig. 3. The confusion matrix for our vision module

These keyframes is ready to be fed into our vision perception module based on the SlowFast Network [18] architecture for visual feature extraction. This sampling approach maximizes the structural characteristics of the SlowFast Network, where the slow pathway is responsible for extracting spatial information, and the fast pathway is tasked with extracting temporal information.

Additionally, every 10-second segment of audio data is processed through speech recognition to generate text, which will be inputted into a language perception module designed on the Bidirectional Encoder Representations from Transformers (BERT) [19] architecture for further audio recognition analysis.

C. *Vision perception Module*

Our vision perception module is based on the SlowFast network [18], a dual-pathway architecture in deep learning, addresses video understanding tasks. It comprises two streams: the slow pathway, designed for spatial feature extraction with fewer frames at high resolution, and the fast pathway, which processes more frames at lower resolution for capturing motion or temporal information. Both paths follow the 3D CNN [20] structure. This innovative design allows the network to efficiently balance between spatial richness and temporal resolution. The slow pathway operates at a lower frame rate to capture detailed spatial features, while the fast pathway works at a higher frame rate, focusing on rapid temporal changes. This combination enables the network to be both computationally efficient and effective in capturing diverse aspects of video data, leading to improved performance in tasks like action recognition. The vision model has been pre-trained 2 rounds on the kinetic-400 dataset[21] and our FOS-II dataset respectively to acquire sufficient prior knowledge. We also modified the final classification layers to generate a vision feature vector for future fusion perception.

D. *Language Perception Module*

Our language perception module is based on the BERT model [19], which leverages the Transformer [22] architecture, specifically its encoder component. Unlike traditional

	Name of Video	Original	Denoised	Original(Eng)	Denoised(Eng)	IS
1	13th_Following Instructions_120-130	엄마가 미안한데 엄마 그. 엄마 때려 보면 안 돼. 막 뛰지 마.	아마. 엄마가 미안한데 엄마, 그 엄마 때리, 보면 안 돼.마케니스 마은.	I'm sorry, mom. You can't hit mom. Don't jump.	Maybe. I'm sorry, mom. Mom, hit that mom. You can't look at her.	C
2	Hospital_Playtime_Physical Activity _110-120	나또. 빼 뺏어 엄마 줘 뺏어. 그렇게 아니 하지 말고 이렇게 전게 주지 말고. 이거 빼면 안 돼.	저 또 빼 뺏어. 엄마 줘 뺏어 그렇게 하지 말고 이고 저게 주지 말고 이거 빼면 안 돼.	Me again. Take it from me. Give it to me. Don't do that. Don't give it to me like this. You can't take it out.	Take it away from me again. Give it to me. Don't do that. Don't give it to me. Don't take it out.	C
3	5th_Playtime_350-360	약속 어겼어, 또 손들어 손들고 있어 두 손 다 열까지 하나 둘 똑바로. 셋 손 들어.	또 어디 갔어? 손 들어 손 들고. 있어 하나 연락하지 하나 둘 똑바로 손 들어	You broke your promise. raise your hands again. One, two, straight. Three, raise your hands.	Where is it? Raise your hand. I have one. You should have contacted me. One, two. Raise your hand	SI-
4	15th_Following Instructions_130-140	여기가 저이요. 은야. 은지야 엄마, 물 좀 갖다 줄래 물.	안녕하세요. 왜 인아야 엄마, 물 좀 갔다 둘이.	Bring here, Eunya, Eunjiya, mom, please bring some water to me.	Hello, why ina, mom, bring some water two.	SI+
5	5th_Following Instructions_340-350	일어서 해보자. 요리 여기.오, 여기 위에 5.	민 해보자, 우리 여기 오, 여기 위에. 그고.	Let's stand up and do. Cook here. Oh, over this 5.	Let's min. we are here oh, over this. That.	VI+
6	6th_Playtime_180-190	뿅 자. 엄마 보내 엄마 줘. 엄마도 아니야, 엄마 줘 공고고.	수 엄마 보내. 엄마 줘. 엄마 줘 엄마.	Okay. Send it to mom. Give it to mom. It's not even your mom. Give it to mom.	Send Sue's mom. Give it to mom. Give it to mom.	VI-
7	12th_Playtime_610-620	안니 마 줄게. 물 좀 켜주세요. 어이구 잘했어요.	아이 몇. 아구 잘했어.	I'll give you a hug. Please turn on the water. Good job.	Oh, my. You did a good job.	P
8	Home_Playing Alone1_320-330	외주이 정마리 아니 장단이 아니 하시 중이라고 왔다니까. 아우 아니야. 장하가. 장난친구 아니야, 아니, 니 핸드폰 없잖아 걔 장.	을. 잔바이 아니야 호중이라고 왔다니까 아니가 자하다. 강난치고. 아, 니는 니 핸드폰 없잖아, 걔 정.	I told you that she's here because she's playing Jung Mari, not Jangdan. Oh, no. Jangha is not a joke. She's not a friend. No, she doesn't have a cell phone.	B. It's not Zanbai. It's Hojung. No, it's not. You don't have your cell phone.	
9	Home_Playing Alone1 430-440	설명을 내라 지금 얘기를 했다고 듣고 싶지 않다고 그런 식으로 얘기를 하더래. 니 설명을 듣고 싶지 않다고. 진짜 말할 필요 없다고 막 이러더래.	성명을 얘기 듣고 싶지 않다고 그런 식으로 얘기를 하더래 니 설명고 듣고 싶지 않다고 진짜 말할 필요 없다고 막 이러더래.	Give me an explanation. He said he didn't want to hear about it. He said he didn't want to hear your explanation. He said there was no need to tell me.	He said that he didn't want to hear the statement, and he said that there was no need to say that he didn't want to hear the explanation.	

Fig. 4. Comparison text data from original and denoised version

unidirectional language models, BERT captures deep bidirectional context by considering both left and right contexts of a word during pre-training phase. This capability allows it to extract detailed statistical properties of language. The model comprises multiple stacked layers with self-attention mechanisms and feed-forward neural networks. BERT's input representation combines word embeddings, facilitating the processing of single or paired text segments while retaining positional information. We modify the final layers to make sure this model can extract the feature from the original text input. The feature vector is directly derived from the CLS token's output, which captures essential context information for the input sequence, making it ready for the next fusion layer.

E. Fusion Perception Module

We have decided to use a straightforward and simple structure for fusion perception because we believe that the previous language and vision feature extraction modules have extracted enough information for IS prediction. Keeping the fusion perception structure simple will benefit us by reducing the amount of computing power and improving efficiency. Firstly, we concatenate the vision feature vector and the language feature vector, and then we design a feed-forward network to process the feature vector and make the final decision on the IS.

F. Model Training

The training of our entire model is planned to be divided into three parts. Initially, we will construct and pre-train the visual perception module, followed by the construction and pre-training of the language perception module, and finally the multi-modal perception aggregation module. For both the visual and language modules, we undertake two phases of pre-training. Our initial pre-training utilizes large-scale public datasets to acquire sufficient prior knowledge. Subsequently, we engage in a second round of pre-training with data from our FOS-II dataset, specifically selecting labels tha t are more readily identifiable by the respective modal models. For instance, during the second pre-training of the visual data, we choose data labeled with terms such as "Physical Negative" or "Positive Contact," which are more easily recognized by visual information. An example would be a video where a child hits their parent (Physical Negative), a behavior that requires visual cues like body language for interpretation. The rationale behind this approach is to reinforce the training of the single-modal perception modules, enabling them to extract their corresponding information more effectively, thereby enhancing the overall performance of the model. After completing the construction and pre-training of the single-modal perception modules, we will establish the final aggregation perception classification layer, which will not undergo any pre-training. The entire network will then be trained using the full dataset of our training data.

III. Preliminary Results

We are actively engaged in the process of training our model. The construction and the two-phase pre-training of our vision module have been completed, along with the initial

Hyperparameters / Metrics	θ=0.33, Unbalanced	θ=0.30, Balanced	θ=0.50, Balanced
Accuracy	0.74	0.57	0.54
Precision	0.33	0.43	0.51
Recall	0.69	0.62	0.68
F1-score	0.40	0.50	0.57

TABLE II

SINGLE MODALITY PERFORMANCE METRICS OF OUR VISION MODULE

pre-processing of linguistic data. Our current focus is on the second phase of pre-training for the language perception module. In the forthcoming months, our agenda includes the comprehensive training of the entire model, followed by its systematic evaluation.

A. Experimental Setup

The experiment was executed on a server equipped with four A5000 GPUs, boasting a cumulative graphics memory of 96 gigabytes. Regarding the training of the vision module, the backbone of our model is the SlowFast neural network R50 architecture [18]. A classification layer was integrated to facilitate independent pre-training. The batch size was set at 24. We employed the Adam optimization algorithm[23] in conjunction with multilabel soft margin loss. Additionally, a step learning rate decay strategy, specifically gamma decay, was utilized to mitigate the risk of over-fitting. For the language module, we configured the model with 12 layers transformer encoders and incorporated a classification layer to ensure its capability for independent pre-training.

B. Vision Module Performance

Our visual module, responsible for visual feature extraction, has undergone two rounds of pre-training and has already demonstrated preliminary capabilities in recognizing contextual IS from visual data. However, since the recognition of ISs is a multimodal issue, our visual module has only been tested to recognize the three visual-based ISs (Engaged Activity of Play - EA, Physical Negative - PN, and Positive Contact - C+). The performance in classifying these three ISs is quite good.

1) Metrics for Evaluating Model Performance: We randomly selected 1598 (20%) 10-second segments as the validation set. The remaining 6510 10-second segments, after dropping some data without IS labels, were used for the training set. We chose accuracy, precision, recall, and F1-score as the metrics to evaluate the model, which were computed by:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|} \tag{1}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where:

$$\begin{aligned}\text{Precision} &= \frac{\sum_{i=1}^{N}|Y_i \cap \hat{Y}_i|}{\sum_{i=1}^{N}|\hat{Y}_i|} \\ \text{Recall} &= \frac{\sum_{i=1}^{N}|Y_i \cap \hat{Y}_i|}{\sum_{i=1}^{N}|Y_i|}\end{aligned} \tag{3}$$

2) Addressing Data Imbalance: For the three ISs, the distribution in the dataset is extremely imbalanced. In this study, we used a *random key frames sampling* strategy for data augmentation. Specifically, instead of using a uniform key frames selection method to sample the original videos, we first divided each video into segments equal to the number of key frames needed (32 in our study), and then randomly selected one frame from each segment as a key frame. This approach ensures that even when the same video segment is used for training, the key frames (training data) fed into the neural network are different, thereby achieving data augmentation. We replicated the videos corresponding to the less frequently occurring ISs (minority class) and randomly dropped videos corresponding to the more frequently occurring ISs (majority class), resulting in a relatively balanced dataset.

3) Visual module classification result: Table II and Figure 3 demonstrate the performance of our model on both the original unbalanced dataset and the balanced dataset. The symbol $\boldsymbol{\theta}$ represents the decision threshold in multi-label classification tasks, and it is a hyperparameter during the training process. If the value corresponding to a certain category of IS in the output vector of the model's final layer exceeds this threshold, the model will predict that such IS exists in the video.

While our model trained on the original unbalanced dataset exhibits a relatively high accuracy (74%), it tends to over-predict the majority class, resulting in a relatively low F1 score. Conversely, the model trained on the balanced dataset, although slightly lower in overall accuracy, does not exhibit significant bias, and its F1 score is higher.

Another noteworthy point is that this task involves multi-label classification, where a single video may possess multiple ISs simultaneously. Consequently, a single confusion matrix cannot display the classification results for all categories. Therefore, we have plotted the separate confusion matrix for each IS category.

C. Language Data Pre-processing

To prepare the text-based language data, we segmented each patient's video and audio files into 10-seconds segments. During this process, We extracted both the original

audio files and versions with reduced noise (using Facebook Research's denoiser algorithm [24]). These 10-second audio datasets were then processed through the Return-Zero STT (Speech-To-Text) model, specialized in the Korean language, to extract text data. However, the accuracy of the transcriptions from the noise-reduced audios was lower than from the original audio files. This reduction in accuracy was particularly noticeable when a speaker in the video was conversing with another person; the denoising model tended to filter out the children's voices when the other speaker was closer to the video recorder. In total, 8870 text data entries were collected and categorized by IS between ASD children and their parents/guardians.

Figure 4 presents selected text excerpts from the transcripts of original and denoised audio recordings. The IS for each segment is also indicated. Text 1 is marked with a "C" to denote a complaint from the patient. However, the corresponding audio does not contain discernible speech but rather non-lexical vocal expressions of distress, which were not extracted as text. The acronyms "SI," "VI," and "P" represent "Specific Instruction," "Vague Instruction," and "Praise," respectively. In some instances, determining whether the interaction style is positive or negative is challenging when relying solely on the text. Additionally, videos 8 and 9 depict the patient in solitary play, yet the transcripts include extraneous text from background conversations from parents and other people.

IV. Conclusion and Future Perspectives

This study proposes a multimodal automatic annotation AI model for the FOS-II [6] scale, with a special focus on developing robust algorithms that can work in uncontrolled settings such as *at-home deployments*, utilizing the most advanced architectures based on BERT [19] and SlowFast [18] networks. While this model is currently under development, the completed visual module demonstrates certain capabilities in feature extraction and behavior recognition, offering promising prospects for further research. The language recognition module is in the process of development, with the data preprocessing already completed. We anticipate completing the development of the entire model in the coming months. Upon completion, we plan to conduct testing and evaluation of the model to ensure its efficacy in clinical research in ASD. Moreover, integrating this model with socially assitive robots could enhance its application in real-world scenarios. For example, socially assitive robots equipped with our model can interact with ASD children in home or institution environments, providing real-time feedback and support. These robots can gather sensor data to monitor behaviors, activity patterns, and health status of patients remotely. By linking the robotic sensor data to other digital health system components such as mobile phones, wearables, and smart home devices, we can create a comprehensive IoT ecosystem that offers a broader picture of patient health outcomes. In the longer timeframe, our goal is to develop an intelligent social robotic agent that can provide interactive care to the child in the home environments with enhanced perception. This integration not only aligns with the goals of digital therapeutics but also enhances the overall utility and reach of our model in clinical and home environments.

V. Acknowledgement

All the data collection and labeling process was conducted after receiving internal review approval from the Department of Psychology of Yonsei University, to ensure the privacy of the patients. The data was de-identified for research at GW. Researchers at GW have finished the CITI Human Research, Social & Behavioral Research training and conducted the research under IRB #111540. The study was partially supported by NSF Grant #1846658 ("CAREER: Social Intelligence with Contextual Ambidexterity for Long-Term Human-Robot Interaction and Intervention (LT-HRI2)").

References

[1] F. Chiarotti and A. Venerosi, "Epidemiology of autism spectrum disorders: A review of worldwide prevalence estimates since 2014," *Brain Sci*, vol. 10, no. 5, p. 274, 2020.

[2] World Health Organization, "Autism Spectrum Disorders Fact Sheet," https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders, [Accessed: March 29, 2023].

[3] I. Hertz-Picciotto and L. Delwiche, "The rise in autism and the role of age at diagnosis," *Epidemiology (Cambridge, Mass.)*, vol. 20, no. 1, p. 84, 2009.

[4] 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.

[5] 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.

[6] M. R. Sanders and T. Glynn, "Training parents in behavioral self-management: An analysis of generalization and maintenance," *Journal of Applied Behavior Analysis*, vol. 14, no. 3, pp. 223–237, 1981.

[7] M. Lee and K. Chung, "Development of parent child interaction-direct observation checklist (pci-d) for children with developmental disabilities," *Journal of Rehabilitation Psychology*, vol. 23, no. 2, pp. 367–395, 2016.

[8] B. Xie and C. H. Park, ""can you guess my moves? playing charades with a humanoid robot employing mutual learning with emotional intelligence," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 667–671. [Online]. Available: https://doi.org/10.1145/3568294.3580170

[9] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder—a personalized computational modeling approach," *Frontiers in Robotics and AI*, vol. 7, p. 43, 2020.

[10] H. Javed, R. Burns, M. Jeon, A. M. Howard, and C. H. Park, "A robotic framework to facilitate sensory experiences for children with autism spectrum disorder: A preliminary study," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–26, 2019.

[11] H. Javed and C. H. Park, "Interactions with an empathetic agent: Regulating emotions and improving engagement in autism," *IEEE robotics & automation magazine*, vol. 26, no. 2, pp. 40–48, 2019.

[12] 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.

[13] 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.

[14] P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, and M. A. Armin, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, vol. 9, no. 6, 2023.

[15] N. M. Rad and C. Furlanello, "Applying deep learning to stereotypical motor movement detection in autism spectrum disorders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 1235–1242.

[16] K. Ganesh, S. Umapathy, and P. Thanaraj Krishnan, "Deep learning techniques for automated detection of autism spectrum disorder based on thermal imaging," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 235, no. 10, pp. 1113–1127, 2021.

[17] I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani, "Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques," *Electronics*, vol. 11, no. 4, p. 530, 2022.

[18] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[24] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.