AV-FOS: Transformer-Based Audio-Visual Multimodal Interaction Style Recognition for Children With Autism Using the Revised Family Observation Schedule 3rd Edition (FOS-R-III)

Zhenhao Zhao, *Member, IEEE*, Eunsun Chung, Kyong-Mee Chung, and Chung Hyuk Park, *Member, IEEE*

***Abstract*—Challenging behaviors in children with autism is a serious clinical condition, oftentimes leading to aggression or self-injurious actions. The Revised Family Observation Schedule 3 rd Edition (FOS-R-III) is an intensive and fine-grained scale used to observe and analyze the behaviors of individuals with autism, which facilitates the diagnosis and monitoring of autism severity. Previous AI-based approaches for automated behavior analysis in autism often focused on predicting facial expressions and body movements without generating a clinically meaningful scale, mostly utilizing visual information. In this study, we propose a deep-learning based algorithm with audio-visual multimodal-data clinically coded with the FOS-R-III, named AV-FOS model. Our proposed AV-FOS model leverages transformer-based structure and self-supervised learning to intelligently recognize Interaction Styles (IS) in the FOS-R-III scale from subjects' video recordings. This enables the automatic generation of the FOS-R-III measures with clinically acceptable accuracy. We explore the IS recognition using a multimodal large language model, GPT4V, with prompt engineering provided with FOS-R-III measure definitions as the baseline for this study and compare with other vision-based deep learning algorithms. We believe this research represents a significant advancement in autism research and clinical accessibility. The proposed AV-FOS and our FOS-R-III dataset will serve as a gateway toward the digital health era for future AI models related to autism.**

***Index Terms*—Autism spectrum disorder, challenging behaviors, multimodal learning, self-supervised learning, video understanding.**

I. Introduction

AUTISM Spectrum Disorder (ASD), or autism, is a life-long neuro-developmental condition [1]. The increasing prevalence of ASD among children in the United States has become a significant developmental issue. Over the past decades, the rate has been steadily rising, with 1 in 36 children now diagnosed with autism [2], [3], [4], [5]. Individuals with ASD, or autistic individuals, experience difficulties in communication and social interaction, exhibit restricted interests, and engage in repetitive behaviors. These characteristics impact their daily activities and social functioning across various settings such as school, work, and other areas of life [6], [7]. One of the more clinically important characteristics with autistic individuals is the challenging behaviors (CBs), such as self-injurious behaviors, aggression and disruptive behaviors [8]. These CBs not only hinder social interaction but also frequently result in critical health implications for the individuals themselves or others. Despite their clinical importance, tracking these behaviors in daily settings remains a significant challenge. Currently, monitoring CBs primarily relies on regular clinical evaluations conducted in office settings, which imposes considerable burdens and restrictions on families of autistic individuals. Moreover, this approach is cost-prohibitive and unsuitable for long-term continuous observation. The sporadic nature of certain episodes may further lead to discrepancies between diagnostic outcomes and actual behavioral patterns. Therefore, developing automated tools capable of analyzing the interactive behaviors between autistic children and their caregivers is not only beneficial for the diagnosis and treatment of children but also essential for reducing the burden on caregivers. Additionally, such tools would facilitate long-term monitoring, enabling more accurate diagnoses and a better understanding of behavioral trends over time.

One of the clinical measures that has been established for rigorous and fine-grained coding of children behaviors is the Revised Family Observation Schedule 3 rd edition (FOS-R-III) [9],

Received 28 June 2024; revised 8 January 2025; accepted 5 February 2025. Date of publication 13 February 2025; date of current version 9 September 2025. This work was supported 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)". *(Corresponding author: Chung Hyuk Park.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Department of Psychology at Yonsei University under Application No. Psychology-59.

Zhenhao Zhao and Chung Hyuk Park are with the Department of Biomedical Engineering, School of Engineering and Applied Science, The George Washington University, Washington, DC 20052 USA (e-mail: zzhao98@gwu.edu; chpark@gwu.edu).

Eunsun Chung and Kyong-Mee Chung are with the Department of Psychology, College of Liberal Arts, Yonsei University, Seoul 03722, South Korea (e-mail: eun930320@gmail.com; kmchung@yonsei.ac.kr).

Digital Object Identifier 10.1109/JBHI.2025.3542066

which is a direct observation tool designed to assess parent-child interactions across various contexts. In autism research, FOS-R-III is frequently utilized in both clinical and research settings to identify and evaluate parent-child interactions, particularly in relation to CBs. This tool provides valuable insights for developing interventions and support strategies for autistic children by examining their social contexts and dynamics [10]. Currently, FOS-R-III data is manually encoded by trained observers through video interactions between autistic children and their caregivers, a process that is both time-consuming and labor-intensive. Developing an automated FOS-R-III encoding algorithm suitable for clinical settings could significantly reduce the workload for clinicians and researchers, ultimately benefiting many autistic children and their families.

To design the automated tools to achieve this goal, we apply a multimodal sensor-based approach with artificial intelligence (AI). In recent years, multimodal perception algorithms have found numerous applications in the field of human behavior detection [11], [12]. Transformer-based multimodal models, in particular, have demonstrated strong capabilities across various video understanding tasks [13], [14]. However, these transformer-based models heavily rely on extensive enterprise-level computational resources and large datasets for training, while the clinical observational data of autistic children is typically not easily accessible due to privacy issues and the size of the data is small.

To address these challenges, we first introduce a high-quality FOS-R-III dataset, meticulously annotated by experts. This dataset comprises nearly 25 hours of videos featuring autistic children, with Interaction Styles (IS) from FOS-R-III annotated every 10 seconds. This dataset is highly suitable for both supervised and unsupervised learning in deep learning models, facilitating future research on deep learning algorithms for autistic children. Secondly, we propose a audio-visual transformer-based model (AV-FOS) for recognizing interaction styles in autistic children, which features relatively manageable computational requirements and real-time inference speed. The AV-FOS model was trained and tested on our proposed FOS-R-III dataset. As a baseline, we compared it with the enterprise-level model (GPT-4V [15]) combined with prompt engineering. As comparison models, we applied our dataset to two vision-based behavior understanding AI models SlowFast Networks [16] and vision transformer [17] and conducted an ablation study. Our AV-FOS model exhibited superior performance and inference speed compared to the baseline as well as comparison models.

II. Related Work

A. FOS-R-III Clinical Application and Case Study

The FOS-R-III is a validated coding system designed to capture negative behaviors and interaction styles in children with ASD and their parents at 10-second intervals. It is widely used for assessing challenging behaviors. Sander et al. [18] employed FOS-R-III in a study evaluating behavioral changes before and after a parenting program, finding a significant reduction in negative child behaviors in the intervention group. Similarly, Pasalich et al. [19] used it to examine the impact of callous–unemotional (CU) traits and ASD symptoms on child conduct problems and parent-child interactions, demonstrating its versatility in analyzing behavioral dynamics.

Despite its effectiveness, FOS-R-III coding has primarily relied on manual processes, which are time-consuming and labor-intensive. ASD behavior assessment services also face challenges such as specialist shortages and limited access, imposing financial and time burdens on families [20]. Developing a deep learning model for automated video analysis and real-time assessment could mitigate these issues, enabling early detection of behavioral changes and timely interventions to support affected families.

B. Multimodal Learning for Behavior Recognition

Multimodal behavior recognition is a highly active research field. Previous studies have focused on various aspects, such as emotion/behavior recognition using video and text information [21], [22], [23], [24], [25], or action recognition using various visual modalities like optical flow and skeleton tracking [26], [27], [28], [29], [30]. However, the previous studies have limitations on the modality of inputs, limiting the bandwidth of contextual understanding. Thus, our study focuses on recognizing behaviors of autistic children and their caregivers using audio and video modalities capable of providing fine-grained clinical explainability. While similar studies using audio-visual modalities exist [31], [32], they do not employ self-supervised learning strategies or provide fine-grained clinical information [31], [32], [33].

Noteworthy are the two AI-based multimodal models capable of audio+video understanding: Audio-Visual Masked Autoencoder (AV-MAE) [34] and Contrastive Audio-Visual Masked Autoencoder (CAV-MAE) [35]. We adapt these state-of-the-art transformer-based approaches and provide a customized architecture advancing from AV-MAE and CAV-MAE to self-learn the clinical measures in FOS-R-III scale and provide explainable AI module on audio-visual inputs. Our proposed AV-FOS model adapts similar pre-training algorithms as the CAV-MAE but adds new strategies to achieve supervised learning with fine-grained self-built clinical dataset. Furthermore, given the limited capacity of the CAV-MAE model to perceive visual temporal information, we address this limitation through targeted optimizations.

C. Deep Learning-Based Autism Research

There has been extensive research utilizing deep learning techniques in studies of autistic children. Some studies focus on emotion recognition in autistic patients based on their facial expressions [36], [37], while others utilize visual information to recognize simple actions such as clapping and jumping [38], [39], [40]. Additionally, some studies integrate multimodal data, such as video, audio, electroencephalograms, and eye-tracking information, to extract basic facial and emotional features using deep learning models. These extracted features are then analyzed to facilitate the detection of ASD. [41], [42]. However, these studies have not employed large-scale multimodal self-supervised pretraining strategies. At the same time, the clinical application of recognizing only facial expressions and

simple actions is limited. Implementing deep learning methods to automatically recognize behaviors within a comprehensive clinical schedule can play a highly beneficial role not only in diagnosing autistic patients but also in preventing and treating ASD.

D. Multimodal Prompt Engineering

The advent of large AI models, especially following the release of ChatGPT, has introduced a powerful alternative to traditional model training, demonstrating strong performance across fields like medicine and law [43], [44]. GPT-4 V, OpenAI's most advanced publicly available multimodal model, has shown substantial capabilities in visual understanding and language-vision tasks through prompt engineering [44], [45], [46]. Additionally, research highlights its potential in psychology and autism-related behavior recognition [47], [48].

However, we did not fine-tune GPT-4 V as a benchmark for three reasons: 1) its large parameter count incurs high computational costs without added clinical value, 2) inference GPT4V model is impractical for local hospital deployment due to hardware constraints, and 3) OpenAI has not open-sourced GPT-4 V's weights, and available fine-tuning options on their website do not include GPT4V.

Thus, we selected the GPT-4 V + Prompt Engineering approach as our baseline for the FOS-R-III IS Encoding task.

TABLE I
THE DESIGN OF INSTRUCTION LISTS

	Categories	Tasks
A	A. Gross motor control	Walks 10 steps by himself/herself
	B. Fine motor control	Leaves marks or draws on paper with pencil or crayon
	C. Social interaction	Follows instructions when asked to wave or clap his/her hands
	D. Language comprehension	Follows verbal instructions such as "put it over there" or "bring it over here"
	E. Language usage	Answers simple questions by shaking his/her hand or saying "yes/no"
	F. Table manner	Drinks water from the cup without spilling it
	H. Wearing clothes	Extends his/her limbs when changing clothes
	I. Personal hygiene	Washes his/her hands by the sink with running water
	L. Mathmatical ability	Counts from 1 to 5
	M. Problem solving	Chooses a particular tool/material out of many tools/materials
B	A. Gross motor control	Pours water from the cattle/jar into the cup
	B. Fine motor control	Closes the zipper on his/her clothes when wearing them
	C. Social interaction	Plays simple games (e.g., rolling balls) with other people
	D. Language comprehension	Identifies his/her name from a group of names that incudes at least 4 other names
	E. Language usage	Names familiar objects such as cup, blanket, or ball.
	F. Table manner	Eats food with fork
	G. Personal care	Flushes the toilet after using it
	H. Wearing clothes	Wears shoes (shoes without laces) correctly
	I. Personal hygiene	Uses handkerchief or tissue to blow and wipe his/her nose
	J. Household chores	Disposes trash at appropriate places
C	A. Gross motor control	Catches bouncing ball (e.g., tennis ball) with two hands
	B. Fine motor control	Screws or places small components such as screws in the right place
	C. Social interaction	Searches or remebers his/her friends' phone number and calls them
	D. Language comprehension	Searches for the needed information from dictionary or encyclopedia
	E. Language usage	Write his/her full name correctly with any assistance
	F. Table manner	Uses knife to cut the food into small pieces if it is too large to eat
	H. Wearing clothes	Ties shoelaces so they do not become untied
	I. Personal hygiene	Takes care of his/her nails (e.g., cutting, grinding) when needed
	J. Household chores	Uses dustpan after sweeping the floor with a broom
	M. Problem solving	Asks an appropriate person for a tool or material when in need
D	A. Gross motor control	Does at least 6 push ups
	B. Fine motor control	Folds the letter into thirds, puts it in an evelope and seals the envelope with glue
	C. Social interaction	Plans to invite people into the house
	D. Language comprehension	Understands news articles or books after reading them
	E. Language usage	Summerizes news articles or books after reading them
	F. Table manner	Uses knife to cut the food into small pieces if it is too large to eat
	H. Wearing clothes	Wears innerwear first before wearing clothes
	I. Personal hygiene	Fixes his/her hair in front of the mirror
	J. Household chores	Cleans with a vacuum cleaner
	M. Problem solving	Asks an appropriate person for a tool or material when in need

III. METHODOLOGY

A. Dataset

1) Dataset Description: This dataset was designed to measure fine-grained FOS-R-III scales for detecting challenging behaviors in autistic children. Researchers recorded videos in participants' homes at the invitation of parents, providing realistic data to enhance clinical services such as ASD treatment, severity diagnosis, and symptom management. This real-life setting underscores the dataset's high clinical value.

The dataset comprises 216 videos, each 5 to 15 minutes long, from 83 participants. The videos were recorded at a frame rate of 30 frames per second, and the corresponding audio was captured at a sample rate of 16,000 Hz. Children with ASD were diagnosed by licensed clinicians, while those without a confirmed diagnosis met the ASD screening cutoff (≥ 15) on the Social Communication Questionnaire (SCQ) [49]. Participants had a mean age of 9.72 years (SD = 4.77), with a male-to-female ratio of approximately 7:3.

Children performed daily tasks designed to assess cognitive, motor, and social skills. Current data focus on children aged 1 to 12, though tasks can be adapted for adolescents and adults in future studies. Problem behaviors ranged from mild to severe, evaluated using the Problem Behavior Checklist [50]. This checklist measures 14 common behaviors (e.g., self-injury, aggression, repetitive movements, noncompliance, feeding issues, hyperactivity) on a 5-point Likert scale [51], with total scores ranging from 14 to 70. Higher scores indicate more frequent or severe behaviors, and participants in this study had a mean score of 33.00, reflecting moderate severity.

Handheld cameras were deliberately chosen to simulate uncontrolled environments, as this introduces a level of noise that enhances the model's robustness to real-world scenarios. While advanced IP-based cameras could provide higher resolution and stability, relying on handheld cameras ensures broader applicability by enabling future diagnostic systems to operate effectively without requiring complex and costly recording setups. Each video features one of three tasks: 1) playing with specific toys, 2) following a series of instructions (four versions available, as shown in Table I), or 3) free play.

2) Dataset Annotation: The videos in this dataset are annotated every 10 seconds using the FOS-R-III structured interval-based coding system to capture interaction styles (IS) between children and their caregivers, which serve as labels for training deep learning models. A total of 23 IS types are coded, encompassing both parental IS (e.g., Praise (P), Affection (AF)) and child IS (e.g., Non-compliance (NC), Opposition (O)). Some IS types are marked with positive or negative symbols to indicate emotional tone; for example, SA+ denotes positive social attention, while SA- represents negative social attention. A detailed overview of IS codes is provided in Table II, and Fig. 1 illustrates several examples of IS annotations corresponding to video frames. If a behavior occurred during the interval, it was recorded as "1". Fig. 2 shows the coding sheet used during the annotation process.

The coding process was conducted manually by trained research assistants, who observed video recordings and documented whether a behavior occurred during each 10-second interval. Five trained graduate students from the Department of Psychology of Yonsei University served as human coders

TABLE II
THE EXPLANATION OF EACH IS AND THE CORRESPONDING FREQUENCY IN THE FOS-R-III DATASET

IS Code	IS Name	Frequency
AD	Adhesive Demand	41
AV	Appropriate Verbal Interactions	1464
Aff_child	Children Affection	24
Aff_parent	Parent Affection	329
C+	Positive Contact	2223
C-	Negative Contact	15
CP	Complaint	178
EA	Engaged Activity of Play	3630
Int_child	Children Interrupt	1
Int_parent	Parent Interrupt	1
MI	Multiple Instructions	185
NC	Non-compliance	150
O	Opposition	2511
P	Praise	332
PN	Physical Negative	72
Q+	Positive Question	1586
Q-	Negative Question	4
S+	Positive Social Attention	5086
S-	Negative Social Attention	13
SI+	Positive Specific Instruction	799
SI-	Negative Specific Instruction	13
VI+	Positive Vague Instruction	2983
VI-	Negative Vague Instruction	20

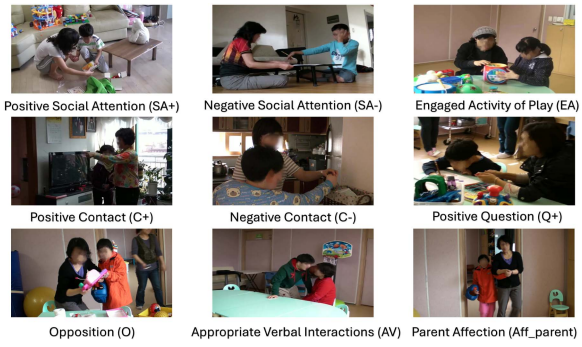


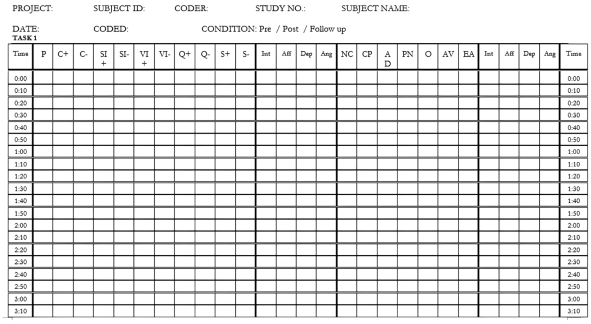
Fig. 1. A subset of the IS examples, accompanied by a single frame from the corresponding videos, is displayed. All images have been anonymized to safeguard the privacy and confidentiality of the participants.

under the supervision of a licensed clinical psychologist with Board Certified Behavior Analyst (BCBA) credentials. Coders underwent extensive training, including 20 hours of practice and evaluation, to ensure annotation accuracy. They worked in pairs to establish inter-observer reliability, and inter-rater reliability was calculated on 30% of the dataset, yielding a 90% agreement rate, exceeding the acceptable threshold of 80% [52].

This rigorous annotation process ensures reliable labels for studying behavior patterns and training machine learning models.

B. Data Preprocessing

For videos originally ranging in length from 5 minutes to 15 minutes, we initially performed a trimming process to establish a dataset comprising clips of 10-second duration each, annotated with corresponding Interaction Styles (10 s FOS-R-III Dataset).


PROJECT: SUBJECT ID: CODER: STUDY NO.: SUBJECT NAME:
DATE: CODED: CONDITION: Pre / Post / Follow up
TASK 1

Time	P	C+	C-	SI+	SI-	VI+	VI-	Q+	Q-	S+	S-	Int	Aff	Dup	Ang	NC	CP	AD	PN	O	AV	EA	Int	Aff	Dup	Ang	Time
0:00																											0:00
0:10																											0:10
0:20																											0:20
0:30																											0:30
0:40																											0:40
0:50																											0:50
1:00																											1:00
1:10																											1:10
1:20																											1:20
1:30																											1:30
1:40																											1:40
1:50																											1:50
2:00																											2:00
2:10																											2:10
2:20																											2:20
2:30																											2:30
2:40																											2:40
2:50																											2:50
3:00																											3:00
3:10																											3:10

Fig. 2. The coding sheet of the annotation.

Subsequently, we utilized the open-source Sound eXchange [53] software and the OpenCV [54] library to extract audio and video information from each 10-second video clip for further processing.

For visual information processing, we adopt three approaches to sample and preprocess 10 s video data, aiming to maximize the preservation of both spatial and temporal information. In all three approaches, the final output consists of 196 visual patches, which are input into the model for attention computation, feature extraction, and IS prediction:

$$\mathbf{v} = [v^1, v^2, \ldots, v^{196}] \tag{1}$$

Approach 1 - Middle Frame Spacial Attention: We select the central frame of the video as the keyframe, resize it to 224× 224 pixels, and divide it into 196 square patches.

Approach 2 - Cross-Frame Attention: The video is divided into four temporal segments, and one keyframe is selected from each. These keyframes are resized to 112× 112 pixels and divided into 49 square patches each, collectively forming 196 patches.

Approach 3 - Averaged Key Frame Attention: We extract one keyframe each from the first, middle, and final thirds of the video, compute their pixel-level average image, resize it to 224× 224 pixels, and divide it into 196 square patches.

Fig. 3 provides a visual comparison of these three approaches.

The first approach prioritizes high-quality spatial information but includes minimal temporal information. The latter two approaches preserve more temporal information by slightly compromising spatial resolution. After evaluation, our Averaged Key Frame Attention demonstrated the best performance; thus, we selected this model for further analysis. Detailed results and discussion can be found in Section IV-F3: Ablation Study - Visual Temporal Information Perception.

For audio processing, the raw waveforms were first normalized by subtracting their mean value, centering the signals and ensuring consistent amplitude across all samples. The audio maintains its native sample rate (16000 HZ), preserving the original quality of the recordings. And then, Mel-filter bank (fbank) features were then extracted using a Hanning window with a window size of 25 ms and a frame shift of 10 ms. The extraction process generated 128-dimensional log Mel-filter bank features for each frame, resulting in a time-frequency representation of the audio data. To ensure uniform input dimensions for the model, the extracted spectrograms were adjusted to a

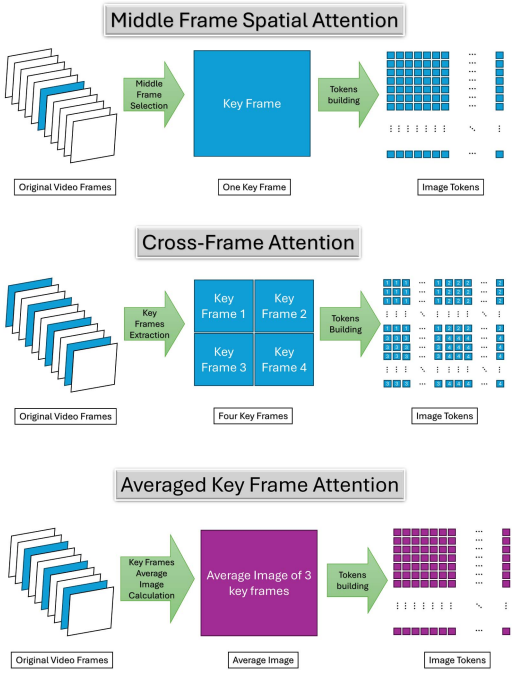


Fig. 3. Comparison of three spatial-temporal attention approaches.

fixed temporal length of 1024 frames through zero-padding for shorter spectrograms or trimming for longer ones.

Finally, the spectrograms were divided into 512 square patches of size 16×16, following a consistent representation format for input into the model. This pre-processing pipeline was designed to preserve critical temporal and spectral information, ensuring that the audio features were robust and aligned with the model architecture:

$$\mathbf{a} = [a^1, a^2, \ldots, a^{512}] \tag{2}$$

C. Transformer-Based Encoder and Decoder

In both the pre-training and formal model structures, the Transformer-based Encoder and Decoder are integral components of our model. Therefore, this section introduces their internal structural details to facilitate the subsequent discussions on the pre-training and formal structures of the model in the following sections.

1) Tokenization: Initially, in the Tokenization phase, we embed not only positional information but also modality information. Specifically, for patch embedding, we use learnable linear projection (LP) layers to process the original square patch $p_m^i \in \mathbf{a} \cup \mathbf{v}$, where each modality $m \in \{\text{audio}, \text{video}\}$ and and i denotes the patch number. In the positional embedding (PE_m^i), a fixed modality-specific 2-D sin-cos embedding strategy is employed. Modality embedding is accomplished using trainable parameters ω. Ultimately, by performing element-wise addition, we obtain the sequence of tokens input into the transformer block. Each token t in this sequence has a length, or embedding dimension, of 768. Consequently, the token t_m^i can be mathematically expressed as:

$$t_m^i = \text{LP}(p_m^i) + \text{PE}_m^i + \omega_m \tag{3}$$

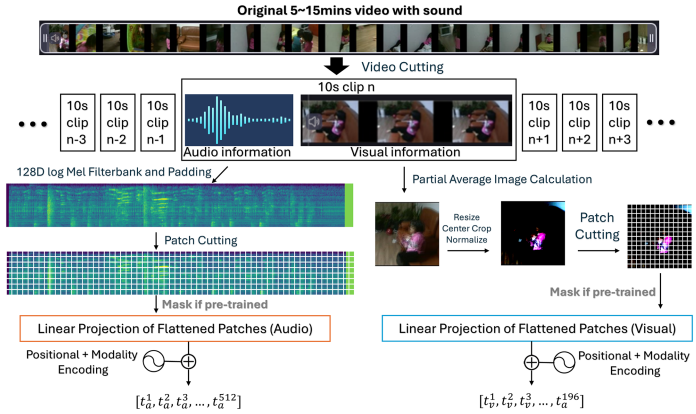


Fig. 4. The data preprocessing and tokenization.

The whole process of data pre-processing and tokenization is shown in Fig. 4

2) Transformer Blocks: In each transformer block of the model, the architecture fundamentally adheres to the standard Transformer structure [55]. A transformer block consists of a stack that follows a specific pattern of a Multi-Head Attention layer (MHA), residual connection layers, a Feed-Forward Neural Network / Multilayer Perceptron layer (MLP), and Layer Normalization layers (LN). For each input token sequence $\mathbf{x} = [t_1, t_2, \ldots, t_n]$ and the corresponding output token sequence $\mathbf{y}$, the mathematical expressions are as follows:

$$\begin{aligned} \mathbf{x}' &= \text{MHA}(\text{LN}_1(\mathbf{x})) + \mathbf{x} \\ \mathbf{y} &= \text{MLP}(\text{LN}_2(\mathbf{x}')) + \mathbf{x}' \end{aligned} \tag{4}$$

Here, LN_1 and LN_2 represent the layer normalization steps applied before the multi-head attention and feed-forward neural network.

3) Encoder and Decoder: The encoder $E_m(\bullet)$ and decoder $D_m(\bullet)$ structures are similar to those in the MAE [56] but accept different modality tokens. The encoder consists of a sequence of transformer blocks applied only to visible, unmasked tokens. Conversely, the decoder is also composed of a sequence of transformer blocks; however, the input to the decoder comprises the full set of tokens, including both masked and unmasked tokens. Each masked token is a shared, learned vector that indicates the presence of a missing patch to be predicted, and both positional embeddings and modality embeddings are added to the tokens. For the different modality encoder and decoder, the structure is the same. We assume that this consistent structure will enhance the performance of modality fusion perception for the multimodal task.

D. Self-Supervised Model Pretraining

Our model used pretraining strategy, leveraging relatively low-cost unlabeled data for prior knowledge acquisition, thereby enabling the use of more data for training in future research, which holds greater potential. We adhere to the original CAV-MAE algorithm for our model initialization and pretraining, as depicted in Fig. 5.

1) Loss Function: Generally, our approach aims to leverage the inherent connections within 1) video information and its

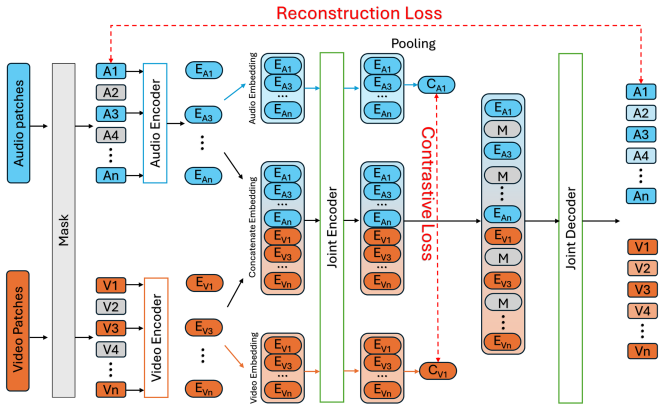


Fig. 5. The pretrained structure of the CAV-MAE. We followed the orginal CAV-MAE paper [35], used reconstruction loss and contrastive loss for the pretrained.

corresponding audio information, and 2) patches within the same contextual data. Consequently, we employ both **Contrastive Loss** and **Reconstruction Loss** as our loss functions. The reduction in contrastive Loss indicates that visual and audio information from the same context are brought closer in the feature space, while data from different contexts are distanced. On the other hand, reconstruction Loss is calculated by initially masking most patches and then generating the masked data using a limited set of features in the feature space along with a Transformer decoder. The loss is then assessed based on the difference between the generated data and the original data. A decrease in Reconstruction Loss indicates that the model has learned the latent connections between contextual data. The computation and application of these two types of losses do not rely on manual annotations, which substantially reduces labeling costs and enhances the model ability to extract features from input data. This is advantageous for our model performance on the self-collected FOS-R-III dataset.

2) Model Structure: The input tokens from the two modalities are initially subjected to a masking process, which obscures 75% of the tokens. Subsequently, these masked tokens are fed into their respective modality-specific encoders, resulting in the preliminary embedding outcomes, denoted as $e^i_{unmask\_a}$ and $e^i_{unmask\_v}i$.

$$e^i_{unmask\_a} = \text{Mask}_{0.75}(E_a(t^i_a))$$
$$e^i_{unmask\_v} = \text{Mask}_{0.75}(E_v(t^i_v)) \tag{5}$$

After passing through the initial unimodality encoders, the two modality embeddings, $e^i_{unmask\_a}$ and $e^i_{unmask\_v}$ are directly input into the Joint Encoder $E_j(\bullet)$ where a Mean Pool operation is conducted to obtain c^{Bi}_a and c^{Bi}_v for computing the contrastive loss. Here, Bi denotes the i-th video clip from the current training batch B. Simultaneously, in order to calculate the reconstruction loss, these two vectors are concatenated and then fed into the Joint Encoder, resulting in the aggregated embeddings sequence $\mathbf{e}_{unmask\_m}$ which is prepared for subsequent reconstruction operations.

$$\mathbf{c}^{Bi}_a = \text{MeanPool}(E_j(\mathbf{e}^{Bi}_{unmask\_a}))$$
$$\mathbf{c}^{Bi}_v = \text{MeanPool}(E_j(\mathbf{e}^{Bi}_{unmask\_v})) \tag{6}$$
$$\mathbf{e}_{unmask\_m} = E_j([\mathbf{e}_{unmask\_a}, \mathbf{e}_{unmask\_v}]) \tag{7}$$

The computation of the contrastive loss $\mathcal{L}_c$ is as follows:

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(s^{Bi,Bi}/\tau\right)}{\sum\limits_{k\neq Bi}\exp\left(s^{Bi,Bk}/\tau\right)+\exp\left(s^{Bi,Bi}/\tau\right)}\right) \tag{8}$$

where $s^{Bi,Bj} = \|c^{Bi}_v\|^T\|c^{Bj}_a\|$ and τ is the temperature.

For the reconstruction loss calculation, we pad $\mathbf{e}_{unmask\_m}$ at the original masked position as $\mathbf{e}_m$ and elementwised add the fixed sinusoidal positional and learnable modality embedding (PE^i_m and ω_m). And then pass the decoder structure to get the reconstruction of the original audio and video patch $\hat{a}_i$ and $\hat{v}_i$.

$$\hat{a^i} = D_j(e^i_a + \text{PE}^i_a + \omega_a)$$
$$\hat{v^i} = D_j(e^i_v + \text{PE}^i_v + \omega_v) \tag{9}$$

We then apply a mean square error reconstruction loss $\mathcal{L}_r$:

$$\mathcal{L}_r = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\sum((\hat{a}^i_{mask}) - \text{norm}(a^i_{mask}))^2}{|a_{mask}|} + \frac{\sum((\hat{v}^i_{mask}) - \text{norm}(v^i_{mask}))^2}{|v_{mask}|}\right] \tag{10}$$

Here, N denotes the mini-batch size, and $|a^i_{\text{mask}}|$ and $|v^i_{\text{mask}}|$ denote the number of masked audio and visual patches, respectively.

Finally, we sum the constractive loss $\mathcal{L}_c$ and reconstruction loss $\mathcal{L}_r$ as the final loss $\mathcal{L}$:

$$\mathcal{L} = \lambda_c\mathcal{L}_c + \mathcal{L}_r \tag{11}$$

Here, $\lambda_c \in [0, 1]$ represents the ratio of the contrastive loss.

3) Model Initialization and Pretrained Dataset: In this study, we utilized the pretrained model weights from the CAV-MAE paper, which were used to initialize our model, specifically CAV-MAE^{scale+}. These weights were obtained through pretraining on the AudioSet dataset [57].

E. FOS-R-III Encoding Model Supervised Learning

To facilitate the model ability to learn more prior knowledge conveniently, during the pretraining phase, we incorporated numerous redundant structures such as decoders and patch masking. However, before proceeding with supervised training on the self-collected FOS-R-III dataset, it is necessary to modify the model structure. This involves removing redundant components while retaining the neural network layers that store the most prior knowledge. Additionally, we introduce appropriate classification layers and employ different loss functions to train the model, optimizing it for the multi-label classification task of FOS-R-III Interaction Styles (IS). This newly constructed and trained network is named the Audio-Visual FOS-R-III Encoding Neural Network (AV-FOS), as illustrated in Fig. 6, which is specifically designed for recognizing FOS-R-III IS in the medical domain.

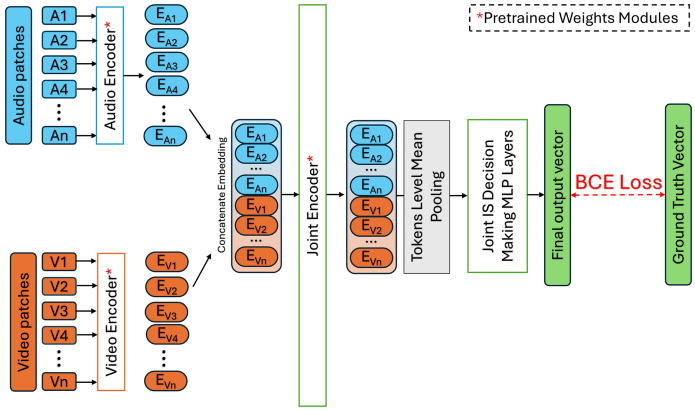


Fig. 6. The proposed FOS-R-III decision neural network: AV-FOS.

The preprocessing and tokenization of input data for AV-FOS remain consistent with previous discussions, except for the elimination of the masking step. The input audio patches $\mathbf{a} = [a^1, a^2, \ldots, a^{512}]$ and video patches $\mathbf{v} = [v^1, v^2, \ldots, v^{196}]$ undergo tokenization and element-wise addition of positional and modality embeddings, resulting in $\mathbf{t}_a = [t_a^1, t_a^2, \ldots, t_a^{512}]$ and $\mathbf{t}_v = [t_v^1, t_v^2, \ldots, t_v^{196}]$, respectively. These tokens are then input into their respective modality-specific encoders, which have been pretrained, followed by concatenation and input into a previously pretrained Joint Encoder to obtain the feature vector $\mathbf{e}_m = [e_a^1, e_a^2, \ldots, e_a^{512}, e_v^1, e_v^2, \ldots, e_v^{196}]$:

$$\mathbf{e}_m = E_j^*[E_a^*(\mathbf{t}_a), E_v^*(\mathbf{t}_v)] \tag{12}$$

Here, the asterisk $(*)$ indicates that the module has undergone pretraining.

Instead of using the traditional class embedding approach for classification [17], we employed a token-level mean pooling strategy: for each embedding dimension (out of 768), we compute the average across all tokens to generate an average token. This average token—a vector of length 768—serves as a mapping of all real-world information in the feature space, which is highly suitable for FOS-R-III classification. This vector is then input into MLP of the decision layer, denoted ISMLP$(\bullet)$, to produce a feature vector $\mathbf{v}_{\text{IS}}$ of length equal to the number of labels (FOS-R-III IS), which is 13:

$$\mathbf{v}_{\text{IS}} = \text{ISMLP}(\text{Mean}(\mathbf{e}_m)) \tag{13}$$

Subsequently, if performing inference, this vector is processed through a Sigmoid function, compared with a manually defined threshold θ, and if it exceeds this threshold, the IS is determined to be present in the input 10-second video:

$$\text{IS}_{\text{detected}} = \{i \mid \text{Sigmoid}(v_{\text{IS}}^i) > \theta\} \tag{14}$$

During the training process, the output of the model, $\mathbf{v}_{\text{IS}}$ is first processed through a Sigmoid function, and then the Binary Cross-Entropy (BCE) Loss $\mathcal{L}_{\text{BCE}}$ is computed with respect to the ground truth one-hot encoded vector $\mathbf{v}_{\text{GT}}$. This loss is then used to guide the training of the model:

$$\mathbf{p}_{\text{IS}} = \text{Sigmoid}(\mathbf{v}_{\text{IS}})$$

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{N}(v_{\text{GT}}^i \cdot \log p_{\text{IS}}^i + (1 - v_{\text{GT}}^i) \cdot \log(1 - p_{\text{IS}}^i)) \tag{15}$$

The currently trained AV-FOS model exhibit 164.512 million parameters.

F. GPT4V Prompt Engineering With FOS-R-III Definitions

We have employed OpenAI state-of-the-art multimodal foundation model, GPT-4 V, [15] combined with prompt engineering as the baseline for our FOS-R-III IS Encoding task.

1) Prompt Engineering: We designed two versions of prompts **(Prompts V1 & V2)**, each consisting of two components: a textual prompt and a visual information prompt. The first version of the visual information prompt **(Prompt V1)** includes the starting, middle, and ending frames from the original 10-second video. The textual prompt guides the model to utilize this three-frame information to facilitate the GPT-4 V in recognizing FOS-R-III IS. The design of the textual component of the first version of the prompt is as follows:

A video is given by providing three frames in chronological order. Please choose one or more appropriate interaction styles or behaviors in the video. Please only reply with the numbers of the interaction styles or behaviors, separated by commas. The candidates of the interaction styles or behaviors are as follows: 1. Appropriate verbal interactions 2. Parent affection 3. Positive contact 4. Complaint 5. Engaged activity of play 6. Multiple instruction 7. Non-compliance 8. Oppositional 9. Praise 10. Positive question 11. Positive social attention 12. Positive specific instruction 13. Positive vague instruction

The second version of the prompt **(Prompt V2)** incorporates a brief explanation of each interaction style within the textual part, while the video component utilizes a method of randomly selecting three key frames. These key frames are extracted randomly from the first third, middle third, and final third of the original 10-second video. The design of the textual prompt for the second version is as follows:

A video is given by providing three frames in chronological order. Please choose one or more appropriate interaction styles or behaviors in the video. Please only reply with the numbers of the interaction styles or behaviors, separated by commas. The candidates of the interaction styles or behaviors are as follows: 1. Appropriate verbal interactions: Appropriate verbal interactions are scored when a child engages in non-aversive, intelligible speech directed at others or self. 2. Parent affection: Parent affection is verbal or non-verbal affection, including words, physical contact, and leveling actions towards the child. 3. Positive contact: Positive contact is friendly, affectionate, or neutral physical interaction initiated or maintained by the parent. 4. Complaint: A complaint is any instance of whining, crying, or other vocal protests displaying temper or discontent. 5. Engaged activity of play: Engaged activity of play is scored when a child quietly plays, observes, or eats without deviance for a full interval. 6. Multiple instruction: Multiple instruction is

scored when a parent gives more than one command or request in a single utterance. 7. Non-compliance: Non-compliance occurs when a child fails to follow a given instruction within five seconds or immediately contradicts it verbally. 8. Oppositional: Oppositional behavior is socially inappropriate or unacceptable child behavior not following specific family rules. 9. Praise: This category scores praise for specific behaviors or characteristics of the child, positively and non-aversively. 10. Positive question: A positive question is a non-aversive, information-seeking utterance from the parent to the child. 11. Positive social attention: Positive social attention is non-aversive verbal or non-verbal engagement by the parent that doesn't fit other categories. 12. Positive specific instruction: A Positive specific instruction is a direct, clear command with a defined behavioral expectation, delivered non-aversively. 13. Positive vague instruction: A Positive vague instruction is an indirect, non-aversive command without a clear behavioral referent.

G. Ethical Considerations and Data Privacy

This study was conducted with strict adherence to ethical and privacy protection guidelines. Data collection and labeling were approved by the Institutional Review Board (IRB) of Yonsei University, Korea, and research conducted in the United States followed the GW IRB #111540. Informed consent was obtained from all participants through an explanatory document approved by the IRB, ensuring they were fully aware of the study's objectives, data usage, and privacy safeguards. Participants provided written consent for video recordings, which were anonymized with unique numeric identifiers to prevent identification. No personally identifiable information (e.g., names, ages, or nationalities) was included in the dataset, and access to the data was strictly limited to IRB-approved researchers. All video data were securely stored in an encrypted, password-protected database and will be permanently deleted upon study completion in compliance with IRB regulations.Additionally, for privacy considerations when using GPT-4 V via the OpenAI API, we consulted OpenAI's official privacy policies [58]. The API ensures that accessed data is deleted within 30 days and is not used for training future AI models. Furthermore, the uploaded data contained no personally identifiable details and did not indicate that the images originated from autistic children. These measures ensured compliance with human subject protection guidelines outlined in the IRB protocol.Researchers interested in using our dataset for academic purposes may contact us directly via email.

IV. Results and Evaluations

A. Experimental Setup

1) Experiment Device: All experiments in this paper, including the training and inference of all deep learning models, were conducted on a server with four NVIDIA A5000 GPUs (Lambda-quad 2). Compared to enterprise-grade servers, this server is not only cost-effective but also moderately sized, akin to a typical household computer, making it highly suitable for deployment in hospital settings.

2) Training Details: During the pre-training and formal training stages, the Encoder part of this model consists of a total of 12 transformer blocks. The single-modality Encoder layer contains 11 transformer blocks, while the joint Encoder comprises only one transformer block. The model Decoder part includes eight transformer blocks. The Transformer blocks in the Encoder have 12 attention heads and an embedding dimension of 768. In contrast, the Transformer blocks in the Decoder have 16 attention heads and an embedding dimension of 512.

TABLE III
The Key Hyperparameters for AV-FOS Model Training Stage

Training stage	Pre-training	Formal Training
Epochs	25	100
Batch size	4×27	128
Initial Backbone LR	2e-4	1e-5
Initial Classification layers LR	-	2e-6
LR decay start epoch	10	5
LR decay rate	0.5	0.95
LR decay step	5	1
Optimizer	Adam weight decay=5e-7 betas=(0.95, 0.999)	

For calculating the contrastive Loss, the temperature τ is set to 0.05, while for computing the CAV-MAE Loss, λ_c is set to 0.01 and the IS decision threshold θ is set to 0.4. The remaining key hyperparameters for both the pre-training and formal training stages on the FOS-R-III dataset are shown in the Table III.

B. Dataset Segmentation

We processed the original dataset into 8,108 ten-second video clips with IS annotations. Table II shows the frequency of each annotation. IS categories with fewer than 100 instances were discarded due to insufficient data for deep learning training. While this reduces the completeness of the Functional Observation Scale (FOS) and impacts immediate clinical applicability, many excluded categories, such as Int_parent, have limited clinical significance. Data collection remains ongoing, and future expansions will allow model retraining to address this limitation. After removing unannotated data, we obtained 8,040 ten-second clips with 13 IS annotation types for training and validation. To enhance clinical relevance and assess generalization to unseen subjects, we employed a subject-based partitioning strategy. Data from 11 subjects formed the validation set (1,867 clips), while the remaining data constituted the training set (6,173 clips). Due to subject-specific behavioral differences, IS distributions vary significantly between training and validation sets, presenting a challenge for the model. Table IV summarizes the IS label distribution in both sets.

C. Metrics for Evaluating Model Performance

In this study, since it is a multi-label task, we evaluated the model using several metrics, including Accuracy, F1 Score, Strict Accuracy, AUC (Area Under the ROC Curve), and mAP. The formulas for these metrics are as follows:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|} \tag{16}$$

TABLE IV
LABEL DISTRIBUTION IN TRAINING AND VALIDATION SETS

	Training		Validation	
Label	Count	Proportion	Count	Proportion
C+	1827	10.91%	396	8.41%
Q+	1210	7.23%	376	7.98%
S+	3821	22.82%	1265	26.86%
AV	883	5.27%	581	12.34%
EA	2880	17.20%	750	15.92%
SI+	672	4.01%	127	2.70%
VI+	2387	14.25%	596	12.65%
O	2053	12.26%	458	9.72%
NC	124	0.74%	26	0.55%
P	287	1.71%	45	0.96%
Aff_parent	288	1.72%	41	0.87%
MI	161	0.96%	24	0.51%
CP	153	0.91%	25	0.53%

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

where:

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

$$\text{Recall} = \frac{\sum_{i=1}^{N} |Y_i \cap \hat{Y}_i|}{\sum_{i=1}^{N} |Y_i|} \tag{18}$$

$$\text{Strict Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(Y_i = \hat{Y}_i) \tag{19}$$

where $\mathbb{I}$ is the indicator function that returns 1 if the argument is true and 0 otherwise.

$$\text{AUC} = \frac{1}{|\mathcal{Y}|} \sum_{k \in \mathcal{Y}} \text{AUC}_k \tag{20}$$

where AUC_k is the AUC for the k-th label.

$$\text{mAP} = \frac{1}{|\mathcal{Y}|} \sum_{k \in \mathcal{Y}} \text{AP}_k \tag{21}$$

where AP_k is the Average Precision for the k-th label. These evaluation metrics reflect not only the absolute performance of the model but also its ability to handle imbalanced datasets.

D. GPT-4 V Result Post-Processing

GPT-4 V generates three types of outputs: ideal outputs, problematic outputs, and unsolvable outputs. Ideal outputs follow the structure specified in the prompt, returning several numerical indices separated by commas. These outputs can be processed with a simple string-splitting algorithm. Problematic outputs return predicted IS but not in the format specified in the prompt, including both numerical indices and IS names. For these cases, we use code to extract the numerical indices. Unsolvable outputs occur when GPT-4 V returns a descriptive statement indicating its inability to process the data. In such cases, the data is manually classified as having no IS present. Table V presents examples of the three different types of outputs.

TABLE V
THE THREE TYPES OF OUTPUTS FOR THE GPT-4 V MODEL

Output	Example	Occurrences (Prompt V1)	Occurrences (Prompt V2)
Ideal Output	2, 3, 5, 11	1863	1845
Problematic Output	5. Engaged activity of play	3	7
Unsolvable Output	The images are too dark to accurately discern any specific interaction styles or behaviors.	1	15

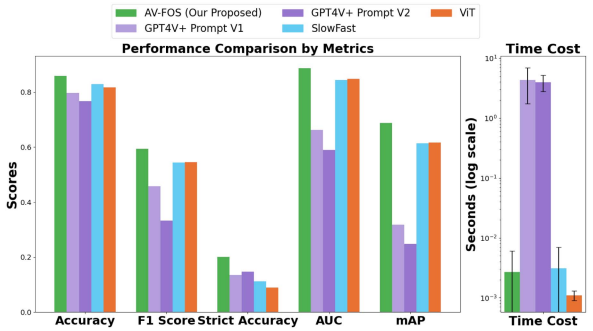


Fig. 7. The performance and time cost comparison.

E. Model Performance

In our experiments, we used GPT-4 V with prompt engineering as the baseline, and also tested two classic models for comparison: the advanced video understanding model SlowFast Networks [16] based on the CNN structure, and the classical visual understanding model vision Transformer (ViT) [17] based on the Transformer structure. Both models were pretrained using supervised learning on large-scale public datasets as mentioned in their original papers. For SlowFast Networks, we selected the R50 architecture, pretrained on the Kinetics-400 dataset [59] and fine-tuned on our FOS-R-III dataset. For ViT, we selected the ViT-base architecture with a patch size of 16x16 for input tokens, pretrained on the ImageNet-21k [60] and ImageNet 2012 [61] datasets, and fine-tuned on our FOS-R-III dataset.

1) General Performance: Table VI and Fig. 7 present a comparison of various performance metrics across different models. The results show that our model significantly outperformed the baseline GPT-4 V model not only in terms of accuracy and the ability to handle imbalanced datasets but also in inference speed. Additionally, our model performance exceeded that of the comparison models, SlowFast Networks and ViT. When tested on subjects that the model had never encountered before, our model still achieved an accuracy of over 85%, demonstrating robust performance. This surpasses the 80% inter-rater reliability standard, though it falls slightly short of the 90% agreement level achieved by human annotators in this study. However, our model has the potential to be further optimized through the continuous collection of new data. Additionally, when faced with an extremely imbalanced dataset, our AUC, mAP, and F1 scores reached 0.88, 0.67, and 0.59, respectively, indicating that the model demonstrates a significant advantage in handling imbalanced datasets. In terms of inference time, the baseline GPT-4 V model lags significantly behind our model. For a 10-second video, our model requires only an average

TABLE VI
THE GLOBAL PERFORMANCE COMPARISON OF DIFFERENT MODELS

	Accuracy	F1 Score	Strict Accuracy	AUC	mAP	Time Cost (Per Sample)
GPT4V+Prompt V1	0.7965	0.4581	0.1355	0.6624	0.3181	4.3349 ± 2.5927
GPT4V+Prompt V2	0.7668	0.3330	0.1468	0.5896	0.2481	3.9792 ± 1.1968
SlowFast	0.8287	0.5437	0.1125	0.8445	0.6138	0.0031 ± 0.0088
ViT	0.8172	0.5448	0.0889	0.8486	0.6167	0.0011 ± 0.0022
AV-FOS (Our Proposed)	**0.8590**	**0.5936**	**0.2003**	**0.8868**	**0.6879**	**0.0018 ± 0.0003**

TABLE VII
DETAILED COMPARISON OF MODEL PERFORMANCE ACROSS CLASSES

	AV-FOS (Proposed)			SlowFast			ViT			GPT4V+Prompt V1			GPT4V+Prompt V2		
	Accuracy	AUC	AP	Accuracy	AUC	AP	Accuracy	AUC	AP	Accuracy	AUC	AP	Accuracy	AUC	AP
AV	0.7156	0.8383	0.6346	0.6920	0.6323	0.4237	0.6883	0.6080	0.4072	0.7070	0.6420	0.4155	0.7070	0.5462	0.3533
Aff	0.9780	0.8178	0.1038	0.9780	0.6198	0.0476	0.9775	0.6859	0.0577	0.7745	0.5867	0.0285	0.5410	0.6342	0.0310
C+	0.7750	0.7559	0.4368	0.7675	0.8003	0.5477	0.7542	0.7795	0.4454	0.7418	0.7716	0.4011	0.6808	0.7762	0.3837
CP	0.9861	0.7010	0.0423	0.9866	0.5627	0.0210	0.9861	0.5420	0.0171	0.9759	0.5143	0.0147	0.9636	0.5081	0.0137
EA	0.6974	0.7624	0.6916	0.5008	0.6010	0.4959	0.4237	0.5577	0.4582	0.4660	0.5375	0.4207	0.5217	0.5672	0.4376
MI	0.9877	0.8273	0.1050	0.9871	0.6422	0.0211	0.9871	0.7767	0.0378	0.9813	0.4970	0.0129	0.9850	0.4989	0.0129
NC	0.9861	0.8154	0.0760	0.9861	0.5969	0.0258	0.9861	0.6509	0.0499	0.9630	0.5073	0.0142	0.9748	0.4943	0.0139
O	0.7788	0.8012	0.5164	0.7392	0.6214	0.3557	0.6818	0.6831	0.3520	0.7525	0.5001	0.2453	0.7542	0.4996	0.2453
P	0.9759	0.7910	0.0747	0.9625	0.6263	0.0674	0.9754	0.6945	0.0588	0.9716	0.5195	0.0304	0.9555	0.4896	0.0241
Q+	0.7986	0.7484	0.3409	0.7970	0.6393	0.2888	0.7991	0.6732	0.3368	0.7981	0.5007	0.2017	0.7949	0.5136	0.2103
S+	0.8366	0.9071	0.9444	0.7761	0.8363	0.9090	0.8024	0.8734	0.9376	0.6631	0.6874	0.7793	0.4879	0.6173	0.7507
SI+	0.9079	0.7606	0.2015	0.9272	0.6404	0.1121	0.9320	0.6543	0.1248	0.8800	0.4940	0.0674	0.9207	0.5086	0.0702
VI+	0.7429	0.8205	0.6604	0.6733	0.7218	0.4925	0.6304	0.6882	0.4386	0.6792	0.4993	0.3190	0.6808	0.5000	0.3192

TABLE VIII
WILCOXON SIGNED-RANK TEST RESULTS BETWEEN AV-FOS AND COMPETING MODELS

Metric	SlowFast		ViT		GPT4V+Prompt V1		GPT4V+Prompt V2	
	W	p-value	W	p-value	W	p-value	W	p-value
Accuracy	7.0	0.0208	9.0	0.0328	0.0	0.0002	5.0	0.0024
AUC	1.0	0.0005	1.0	0.0005	1.0	0.0005	1.0	0.0005
AP	9.0	0.0081	3.0	0.0012	0.0	0.0002	0.0	0.0002

of 0.0018 seconds to complete inference, achieving real-time inference speed. These metrics indicate that our model has high clinical value and can assist doctors and healthcare providers in the diagnosis and risk behavior assessment for autistic children.

2) Class-Wise Evaluation and Error Analysis: The class-wise metrics (Table VII) and confusion matrix (Fig. 8) highlight our model's superior recognition of interaction styles (IS) compared to the GPT-4V+Prompt baseline. Unlike traditional video models like ViT and SlowFast, our model processes both visual and audio inputs, providing a distinct advantage in recognizing IS requiring audio comprehension, such as VI+ and SI+.

Interestingly, visual-only models exhibit some recognition ability for audio-reliant IS due to visual cues like lip movements and head turns. However, our model consistently outperforms them, even in visually dominant IS like Engaged Activity of Play (EA). Wilcoxon signed-rank test results (Table VIII) confirm the statistical significance of our model's superiority across all metrics (Accuracy, AUC, AP), with p-values below 0.05.

Despite its strong performance, our model faces challenges in recognizing minority-class IS, such as Complaint (CP), Parent Affection (Aff_parent), and Non-compliance (NC), due to severe class imbalance. Majority classes like EA (26.86%) and S+ (15.92%) vastly outnumber CP (0.51%) and NC (0.55%), leading to conservative predictions. While this issue affects all models, ours still surpasses SlowFast and ViT in minority-class performance.

To mitigate this, we plan to expand the dataset to improve minority-class recognition. Overall, our model demonstrates robust performance and resilience to data imbalance, outperforming mainstream models and the baseline, with ongoing efforts to enhance its capabilities.

F. Ablation Study

1) Uni-Modal Recognition Performance: To evaluate the effectiveness of the multimodal structure of the AV-FOS model, we decided to conduct ablation experiments by retraining and inferring the model with a single modality. For the single-visual-modality model (V-FOS), we removed the audio input and related processing modules, retaining only the visual tokens for the Joint Encoder. Similarly, for the single-audio-modality model (A-FOS), we removed the visual input while keeping the audio tokens. Both models retained the Joint IS Decision Making Layers and other components, and all pretrained modules underwent the same pretraining as in the AV-FOS model. Training parameters and datasets remained consistent for a fair comparison.

As shown in Table IX, A-FOS outperformed V-FOS, reflecting our task's reliance on audio cues, particularly in categories like VI+ (Positive Vague Instruction) and SI+ (Positive Specific Instruction). Even for visually relevant instances like EA (Engaged Activity of Play), audio signals played a role. However, incorporating visual information further improved performance, with the AV-FOS model achieving the highest accuracy. Both single-modality models performed worse than AV-FOS, with a notable F1 score drop, indicating weaker handling of data imbalance. This highlights the superiority of multimodal perception, which enhances robustness and accuracy. In studying

TABLE IX
THE ABLATION STUDY RESULT

Strategy	Accuracy	F1 Score	Strict Accuracy	AUC	mAP	Time Cost (Per Sample)
A-FOS (Audio)	0.8523	0.5736	0.1912	0.8722	0.6542	0.0015 ± 0.0003
V-FOS (Visual)	0.8226	0.4917	0.1152	0.8296	0.5617	0.0009 ± 0.0003
Without Pretrain	0.8322	0.5328	0.1382	0.8463	0.5630	0.0018 ± 0.0004
Frame Aggregation	0.8544	0.5853	0.1987	0.8881	0.6833	0.0055 ± 0.0015
Cross-Frame Attention	0.8407	0.5455	0.1521	0.8561	0.6879	0.0018 ± 0.0003
Middle Frame Saptial Attention	0.8517	0.5767	0.1918	0.8853	0.6749	0.0018 ± 0.0003
Averaged Key Frame Attention	**0.8590**	**0.5936**	**0.2003**	**0.8868**	**0.6879**	**0.0018 ± 0.0003**

Fig. 8. The confusion matrix for different algorithms.

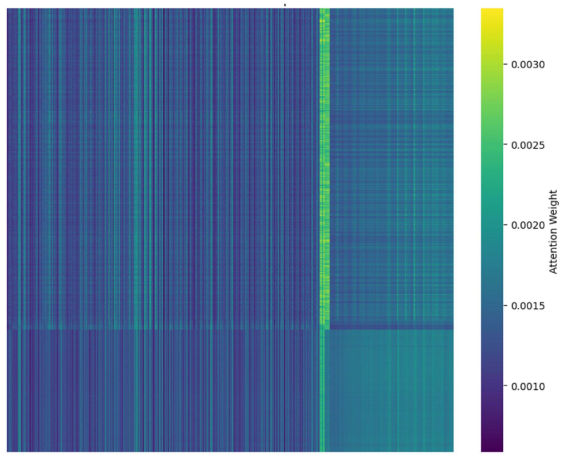


Fig. 9. The attention map for the joint perception layer.

autistic children's behavior, integrating speech, environmental interactions, and facial expressions is crucial, making multimodal models more effective in clinical settings.

2) Without CAV-MAE Pretraining Performance: To assess the impact of CAV-MAE pretraining on the AV-FOS model, we conducted an ablation experiment where all model parameters were randomly initialized, while training methods, hyperparameters, and dataset partitioning remained identical to the original pretrained AV-FOS model. The results, shown in Table IX, indicate that even without pretraining, the AV-FOS model outperformed the GPT-4V+Prompt baseline, achieving over 83% accuracy, demonstrating the strong performance of its multimodal structure. However, the pretrained AV-FOS model still performed better. While accuracy decreased by only 2%, the F1 score and mAP dropped by 6% and 12%, respectively, highlighting the importance of pretraining in handling data imbalance. These findings suggest that pretraining significantly enhances robustness and accuracy, making the model more effective for clinical applications.

3) Visual Temporal Information Perception Module: To enhance the model's ability to perceive visual temporal information, we proposed Cross-Frame Attention and Averaged Key Frame Attention strategies. Experimental results show that Averaged Key Frame Attention outperforms both Middle Frame Spatial Attention and Cross-Frame Attention. This aligns with expectations, as the CAV-MAE pretraining framework extracts single frames rather than multiple frames, limiting Cross-Frame Attention's ability to leverage pretraining knowledge. As a result, its performance is even lower than Middle Frame Spatial Attention. In contrast, Averaged Key Frame Attention retains frame dimensions while averaging pixel values across frames, effectively preserving pretraining knowledge and some spatiotemporal information, leading to superior performance. Compared to Frame Aggregation, a method from CAV-MAE that infers three times using different frames and averages the results, Averaged Key Frame Attention achieves better performance in most metrics while requiring only one inference step, reducing inference time to one-third of Frame Aggregation. This efficiency is crucial for real-world deployment, particularly in clinical settings where computational resources may be limited. Thus, Averaged Key Frame Attention balances computational

feasibility and performance, making it the optimal choice for our model.

G. Inference Visualization

Fig. 9 illustrates the attention distribution within the fusion perception layer. The visualization reveals four distinct attention regions corresponding to: 1) visual-to-visual, 2) visual-to-audio, 3) audio-to-visual, and 4) audio-to-audio. These patterns indicate the model's ability to distinguish attention focus based on semantic relationships across modalities. Notably, strong cross-modal attention reflects effective integration and inter-modal modeling, while significant intra-modal attention underscores the model's robustness in capturing modality-specific features. Overall, these characteristics demonstrate the model's strong capacity for multimodal perception and fusion tasks.

V. Conclusion

To address the challenges in recognizing the complex behaviors and interactions of autistic children, thereby aiding in their diagnosis, symptom assessment/mitigation, and treatment, this study has: 1. Proposed a dataset based on the FOS behavior scale specifically for children with autism. This dataset was constructed from clinically collected data annotated by professionals with medical expertise. 2. Introduced a transformer-based deep learning model, AV-FOS, capable of automatically generating FOS-R-III scales from videos, which holds significant clinical value. This model can utilize self-supervised learning methods to pretrain on large-scale unlabelled video datasets unrelated to autism and make final FOS IS judgments based on both audio and video modalities, demonstrating high accuracy and robustness against imbalanced data. 3. Explored the application of large AI models and prompt engineering in the field of autism behavior recognition.

References

[1] M. M. Hughes et al., "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. [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, 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. 28, no. 10, 2637–2646, 2024. [Online]. Available: https://doi.org/10.1177/13623613241236112

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

[5] M. J. Maenner, 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. SS-2, pp. 1–14, 2023.

[6] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, vol. 4, Washington, DC, USA: American Psychiatric Association, 1994.

[7] National Institute of Mental Health, "Autism spectrum disorder," 2024. [Online]. Available: https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd

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

[9] M. Sanders, L. Waugh, L. Tully, and K. Hynes, "The revised family observation schedule: Fos-r-iii," *Unpublished Manual*, 1996.

[10] M. Lee and K. Chung, "Development of parent child interaction-direct observation checklist (PCI-D) for children with developmental disabilities," *J. Rehabil. Psychol.*, vol. 23, no. 2, pp. 367–395, 2016.

[11] L. N. Nguyen et al., "Non-contact multimodal indoor human monitoring systems: A survey," *Inf. Fusion*, vol. 110, 2024, Art. no. 102457. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524002355

[12] C. A. Casado, M. L. Canellas, and M. B. Lopez, "Depression recognition using remote photoplethysmography from facial videos," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3305–3316, Oct.–Dec. 2023.

[13] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," in *Proc. Int. Conf. Mach. Learn.*, 2021, vol. 2, no. 3, pp. 813–824.

[14] R. Karim and R. P. Wildes, "Understanding video transformers for segmentation: A survey of application and interpretability," 2023, *arXiv:2310.12296*.

[15] OpenAI, "GPT-4V system card," 2023. [Online]. Available: https://openai.com/index/gpt-4v-system-card/

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6201–6210.

[17] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2020.

[18] M. R. Sanders, C. Markie-Dadds, L. A. Tully, and W. Bor, "The triple p-positive parenting program: A comparison of enhanced, standard, and self-directed behavioral family intervention for parents of children with early onset conduct problems," *J. Consulting Clin. Psychol.*, vol. 68, no. 4, 2000, Art. no. 624.

[19] D. S. Pasalich, M. R. Dadds, and D. J. Hawes, "Cognitive and affective empathy in children with conduct problems: Additive and interactive effects of callous-unemotional traits and autism spectrum disorders symptoms," *Psychiatry Res.*, vol. 219, no. 3, pp. 625–630, 2014. [Online]. Available: https://doi.org/10.1016/j.psychres.2014.06.025

[20] 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–e529, 2014.

[21] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 9694–9705.

[22] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 5583–5594.

[23] H. Bao et al., "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2022, vol. 35, pp. 32897–32912. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/d46662aa53e78a62afd980a29e0c37ed-Paper-Conference.pdf

[24] R. Qian, Y. Li, Z. Xu, M.-H. Yang, S. Belongie, and Y. Cui, "Multimodal open-vocabulary video classification via pre-trained vision and language models," 2022,&*arXiv:2207.07646*.

[25] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.

[26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 568–576.

[27] W. Hao et al., "TSML: A new pig behavior recognition method based on two-stream mutual learning network," *Sensors*, vol. 23, no. 11, 2023, Art. no. 5092. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5092

[28] A. Tran and L.-F. Cheong, "Two-stream flow-guided convolutional attention networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3110–3119.

[29] Q. Hu and H. Liu, "Multi-modal enhancement transformer network for skeleton-based human interaction recognition," *Biomimetics*, vol. 9, no. 3, 2024, Art. no. 123.

[30] X. Zhu, Y. Zhu, H. Wang, H. Wen, Y. Yan, and P. Liu, "Skeleton sequence and RGB frame based multi-modality feature fusion network for action recognition," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 18, no. 3, pp. 1–24, 2022.

[31] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, 2021, Art. no. 4913. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4913

[32] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, no. 3, pp. 1325–1341, 2019. [Online]. Available: https://doi.org/10.1007/s11280-018-0548-3

[33] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Adv. neural Inf. Process. Syst.*, 2021, vol. 34, pp. 14200–14213.

[34] M.-I. Georgescu, E. Fonseca, R. T. Ionescu, M. Lucic, C. Schmid, and A. Arnab, "Audiovisual masked autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16144–16154.

[35] Y. Gong et al., "Contrastive audio-visual masked autoencoder," in *Proc. 2023 Int. Conf. Learn. Representations*, 2023.

[36] S. Weigelt, K. Koldewyn, and N. Kanwisher, "Face identity recognition in autism spectrum disorders: A review of behavioral studies," *Neurosci. Biobehavioral Rev.*, vol. 36, no. 3, pp. 1060–1084, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763411002156

[37] J. W. Griffin et al., "Investigating the face inversion effect in autism across behavioral and neural measures of face processing: A systematic review and Bayesian meta-analysis," *JAMA Psychiatry*, vol. 80, no. 10, pp. 1026–1036, 2023. [Online]. Available: https://doi.org/10.1001/jamapsychiatry.2023.2105

[38] A. Ali, F. F. Negin, F. F. Bremond, and S. Thümmler, "Video-based behavior understanding of children for objective diagnosis of autism," in *Proc. VISAPP - 17th Int. Conf. Comput. Vis. Theory Appl.*, Online, France, Feb. 2022, pp. 475–484.

[39] K. Vyas et al., "Recognition of atypical behavior in autism diagnosis from video using pose estimation over time," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process.*, 2019, pp. 1–6.

[40] M. Ruan, X. Yu, N. Zhang, C. Hu, S. Wang, and X. Li, "Video-based contrastive learning on decision trees: From action recognition to autism diagnosis," in *Proc. 14th ACM Multimedia Syst. Conf.*, 2023, pp. 289–300. [Online]. Available: https://doi.org/10.1145/3587819.3590988

[41] M. Cheng et al., "Computer-aided autism spectrum disorder diagnosis with behavior signal processing," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2982–3000, Oct.–Dec. 2023.

[42] J. Han, G. Jiang, G. Ouyang, and X. Li, "A multimodal approach for identifying autism spectrum disorders in children," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2003–2011, 2022.

[43] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023,&*arXiv:2303.13375*.

[44] J. Achiam et al., "GPT-4 technical report," 2023,&*arXiv:2303.08774*.

[45] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae, and T. Hayakawa, "Prompt engineering of GPT-4 for chemical research: What can/cannot be done," *Sci. Technol. Adv. Materials: Methods*, vol. 3, no. 1, 2023, Art. no. 2260300. [Online]. Available: https://doi.org/10.1080/27660400.2023.2260300

[46] B. Meskó, "Prompt engineering as an important emerging skill for medical professionals: Tutorial," *J. Med. Internet Res.*, vol. 25, 2023, Art. no. e50638.

[47] Z. Lian et al., "Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition," *Inf. Fusion*, vol. 108, 2024, Art. no. 102367.

[48] S. Dhingra, M. Singh, S.B. Vaisakh, N. Malviya, and S. S. Gill, "Mind meets machine: Unravelling GPT-4's cognitive psychology," *BenchCouncil Trans. Benchmarks, Standards Evaluations*, vol. 3, no. 3, 2023, Art. no. 100139. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S277248592300056X

[49] S. CHANDLER et al., "Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders," *J. Amer. Acad. Child Adolesc. Psychiatry*, vol. 46, no. 10, pp. 1324–1332, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0890856709618517

[50] N. Y. Shin, "Predictors of parenting stress in mothers of children with developmental disabilities: Demographic, child, parent variables," Master's thesis, Yonsei University, Seoul, South Korea, 2009.

[51] I. E. Allen and C. A. Seaman, "Likert scales and data analyses," *Qual. Prog.*, vol. 40, no. 7, pp. 64–65, 2007.

[52] J. Cooper, *Applied Behavior Analysis*. Upper Saddle River, NJ, USA: Pearson/Merrill-Prentice Hall, 2007.

[53] L. Norskog, C. Bagwell, and S. Contributors, *SoX: Sound eXchange, the swiss army knife of audio manipulation*, 14th ed., 2015. [Online]. Available: http://sox.sourceforge.net

[54] G. Bradski, "The OpenCV library," *Dr Dobb's J. Softw. Tools Professional Programmer*, vol. 25, pp. 120–123, 2000.

[55] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[56] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 16000–16009.

[57] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 776–780.

[58] OpenAI, "Enterprise privacy," 2024. [Online]. Available: https://openai.com/enterprise-privacy/

[59] W. Kay et al., "The kinetics human action video dataset," 2017,&*arXiv:1705.06950*.

[60] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21 k pretraining for the masses," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021.

[61] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.