MMASD: A Multimodal Dataset for Autism Intervention Analysis

Jicheng Li
University of Delaware
Newark, DE, United States
lijichen@udel.edu

Vuthea Chheang
University of Delaware
Newark, DE, United States
vuthea@udel.edu

Pinar Kullu
University of Delaware
Newark, DE, United States
pkullu@udel.edu

Eli Brignac
University of Delaware
Newark, DE, United States
ebrignac@udel.edu

Zhang Guo
University of Delaware
Newark, DE, United States
guozhang@udel.edu

Kenneth E. Barner
University of Delaware
Newark, DE, United States
barner@udel.edu

Anjana Bhat
University of Delaware
Newark, DE, United States
abhat@udel.edu

Roghayeh Leila Barmaki*
University of Delaware
Newark, DE, United States
rlb@udel.edu

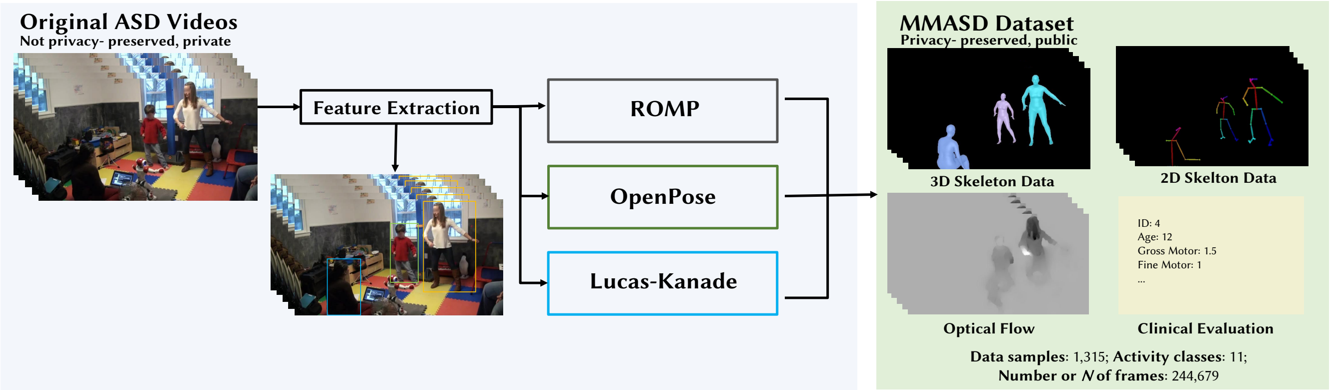

Figure 1: MMASD provides multiple multimodal privacy-preserving features derived from original videos via ROMP [40], OpenPose [5], and Lucas-Kanade [29], including optical flow, 2D / 3D skeleton, and clinical evaluation results. The dataset is publicly accessible for addressing research questions centered on social and behavioral interactions of children with ASD in playful group activities.

ABSTRACT

Autism spectrum disorder (ASD) is a developmental disorder characterized by significant impairments in social communication and difficulties perceiving and presenting communication signals. Machine learning techniques have been widely used to facilitate autism studies and assessments. However, computational models are primarily concentrated on very specific analysis and validated on private, non-public datasets in the autism community, which limits comparisons across models due to privacy-preserving data-sharing complications. This work presents a novel open source privacy-preserving dataset, **MMASD** as a **M**ulti**M**odal **ASD** benchmark dataset, collected from play therapy interventions for children with autism. The MMASD includes data from **32** children with ASD, and **1,315** data samples segmented from more than **100** hours of intervention recordings. To promote the privacy of children while offering public access, each sample consists of four privacy-preserving modalities, some of which are derived from original videos: **(1)** optical flow, **(2)** 2D skeleton, **(3)** 3D skeleton, and **(4)** clinician ASD evaluation scores of children. MMASD aims to assist researchers and therapists in understanding children's cognitive status, monitoring their progress during therapy, and customizing the treatment plan accordingly. It also inspires downstream social tasks such as action quality assessment and interpersonal synchrony estimation. The dataset is publicly accessible via the MMASD project website.

[*] Correspondence: Roghayeh Leila Barmaki (rlb@udel.edu)

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614117

CCS CONCEPTS

• Social and professional topics → People with disabilities; **• Human-centered computing**; **• Computing methodologies → Activity recognition and understanding**; **Machine learning**;

KEYWORDS

multimodal dataset; machine learning; deep learning; autism spectrum disorder; human activity recognition; 2D/ 3D skeleton; privacy-preserving data sharing.

ACM Reference Format:
Jicheng Li, Vuthea Chheang, Pinar Kullu, Eli Brignac, Zhang Guo, Kenneth E. Barner, Anjana Bhat, and Roghayeh Leila Barmaki. 2023. MMASD: A Multimodal Dataset for Autism Intervention Analysis. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614117

1 INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by significant impairments in social communication, as well as difficulties in perceiving and expressing communication cues. Approximately 1 in 54 children are on the spectrum in the United States, resulting in over 1 million affected individuals nationwide [1]. Primary treatment of ASD includes behavioral and psychosocial interventions accompanied by prescribed medications. Behavioral and psychosocial interventions facilitate social and communication development, while medication helps control associated symptoms and comorbid problems [4]. Specifically, psychosocial interventions are diverse in content and can vary in curriculum, structure, discipline, and theme. Prevailing therapeutic interventions include applied behavior analysis and robot-assisted therapy, both of which can provide valuable data for analyzing children's mental development and developing individualized treatment plans.

Various studies have widely used machine learning techniques to facilitate autism research [7, 11, 17, 26, 27, 30, 32]. Compared to traditional methods that rely heavily on human expertise and experience, machine learning approaches can help reduce the need for human labor and associated costs while still achieving decent performance. The autism community has benefited from machine learning techniques in many areas, including but not limited to autism diagnosis [7, 42], emotion recognition [25, 30], and movement pattern assessment [24, 26, 27, 32].

In machine learning research, it is widely accepted to follow a research pipeline that involves developing, applying, and comparing models across multiple benchmark datasets to ensure fair comparisons in performance. However, in the autism community, commonly recognized benchmarks, especially for behavior analysis and activity understanding, are limited due to privacy concerns. Typically, models are validated on a private dataset, and data-sharing roadblocks can restrict the comparison between models. In this sense, the availability of publicly accessible datasets is a crucial first step for the autism community since it allows cutting-edge machine-learning techniques to be trained and validated on ASD datasets. Although some studies have already been conducted, there is still a significant lack of *publicly available, multimodal* datasets that can be used to analyze the *full-body movements* of children during therapeutic interventions.

To overcome some of these ASD data-sharing challenges, we propose a publicly available multimodal ASD dataset, **MMASD**[1]. MMASD maintains privacy while retaining essential movement features by providing optical flow, 2D and 3D skeletons that are derived from the original play therapy videos, thereby avoiding the exposure of sensitive and identifiable raw video footage. Additionally, it includes clinician evaluation results of each child, such as motor function scores. We also provide the intervention activity class labels for overall scene understanding. Overall, MMASD can be used to help therapists and researchers to understand children's cognitive status, track development progress in therapy, and guide the treatment plan accordingly. It also provides inspiration for downstream tasks, such as activity recognition [32], action quality assessment [24], and interpersonal synchrony estimation [27]. In contrast to current datasets, e.g., listed in Table 1, our dataset stands out for the following reasons:

- It is a **publicly accessible** benchmark dataset for movement and behavior analysis during therapeutic interventions featured by diverse scenes, group activities, and participants.
- It includes **multimodal** features such as optical flow, 2D/3D skeletons, demographic, and clinical evaluation data. These features provide **privacy-preserving** approaches to maintain critical and full-body motion information.
- Each scene depicts the same activity performed by a child and one or more therapists, providing an in-place template for comparing typically developing individuals with children with ASD.

The rest of the paper is organized as follows. Further background on relevant autism datasets is presented in Section 2, followed by a presentation of data collection approach in Section 3. Details of the dataset, including statistics, data processing and annotation, are provided in Section 4. Finally, discussion and current limitations of MMASD are described in Section 5 and conclusion in Section 6.

2 RELATED WORK

ASD is characterized by atypical movement patterns, such as repetitive movements, clumsiness, and difficulties with coordination. A deeper understanding of these movement patterns and their association with ASD can aid therapists, clinicians, and researchers in developing more effective interventions and therapies. To this end, several datasets have been developed for movement analysis of children with ASD. These datasets typically involve collecting motion capture or sensor data from children while they perform various activities or specific tasks. The collected data is then analyzed to identify differences and patterns in movement between children with ASD and typically developing children. In this section, we present an overview of existing datasets that focus on movement and behavior analysis of children with ASD.

Movement analysis with humanoid robot interactions datasets. Marinoiu *et al.* [30] presented a dataset and system that use a humanoid robot to interact with children with ASD and monitor their body movements, facial expressions, and emotional states. The

[1] The MMASD Dataset is accessible via https://sites.udel.edu/hci-lab/mmasd-project/.

Table 1: A comparison of related benchmarks and our dataset focused on behavior and movement analysis of children with ASD, including the target population, research focus, data modalities, computational models, data sample, and availability.

Dataset	Target Population	Research Focus	Data Modalities	Computational Models	*N* of Data Samples (Data length)*	Availability
Billing *et al.* [4]	61 ASD	Behavior analysis	Body motion, head pose, eye gaze	SVM, face detector [41], Microsoft Kinect SDK	3,121 sessions (306 h)	Public
Rajagopalan *et al.* [33]	-	Self-stimulatory behaviors detection	Video, audio	Space time interest point (STIP) [22]	75 videos (90 s/video)	Public
Rehg *et al.* [34]	121 total	Behavior analysis	Video, audio, physiological data	SVM, Omron OKAO vision library	160 sessions (3–5 m/video)	On request
DE-ENIGMA [35]	128 ASD	Behavior analysis	Facial landmark, body postures, video, audio	Computer vision, Microsoft Kinect SDK	50 annotated recordings (154 h)	On request
Zunino *et al.* [45]	20 ASD, 20 TD**	Grasping actions	Hand & arm trajectories	LSTM, CNN, Histogram of optical flow (HOG), VLAD [19]	1,837 videos (83 frames/video)	On request
Pandey *et al.* [32]	37 ASD	Behavior analysis	Video, 2D pose, optical flow	Guided weak supervision, Temporal segment networks, Inflated 3D CNN	1,481 video clips	On request
Del Coco *et al.* [12]	8 ASD	Behavior analysis	Facial landmark, gaze, head pose	Constrained Local Neural Fields [3]	6 videos (joint activities), 4 videos (imitative play)	Private
Dawson *et al.* [10]	22 ASD, 82 TD	Phenotyping, head movement	Head pose, facial landmark	Model-based object pose, Computer vision analysis (CVA) [21]	10 videos (1–3 s/video, 49 facial landmarks)	Private
Martin *et al.* [31]	21 ASD, 21 TD	Head movement	Head pose, facial landmark	Computer-vision head tracking (Zface) [20]	252 videos (6 m/video)	Private
MMASD (Ours)	32 ASD	Movement & social behavior analysis	Optical flow, 2D pose, 3D pose, Clinical score	ROMP [40], OpenPose [5], Lucas-Kanade [29]	1,315 videos (7 s & 186 frames/video)	Public

**Certain works provided information on the duration of individual videos, whereas others presented the overall length of the dataset.*
***TD: Typically Developing.*

results show that children significantly improved their ability to recognize emotions, maintain eye contact, and respond appropriately to social cues. They identified several effective techniques for detecting and analyzing the children's emotional and behavioral responses, such as analyzing the frequency and duration of specific behaviors. Billing *et al.* [4] proposed a dataset of behavioral data recorded from 61 children with ASD during a large-scale evaluation of robot-enhanced therapy. The dataset comprises sessions where children interacted with a robot under the guidance of a therapist and sessions where children interacted one-on-one with a therapist. For each session, they used three RGB and two RGBD (Kinect) cameras to provide detailed information, including body motion, head position and orientation, and eye gaze of children's behavior during therapy.

Another dataset related to humanoid robot interactions with ASD children is DE-ENIGMA [35], which includes using a multimodal human-robot interaction system to teach and expand social imagination among children with ASD. The DE-ENIGMA dataset comprises behavioral features such as facial mapping coordinates, visual and auditory, and facilitates communication and social interaction between the children and the robot. The authors indicated that the DE-ENIGMA could be used as an effective tool for teaching and expanding social imagination in children with ASD. They also suggest that the usage of a multimodal human-robot interaction could be a promising approach for developing interventions for children with ASD that aim to improve their social skills and promote better social integration.

Eye movement and vocalization datasets. Duan *et al.* [13] introduced a dataset of eye movement collected from children with ASD. The dataset includes 300 natural scene images and eye movement data from 14 children with ASD and 14 healthy individuals. It was created to facilitate research on the relationship between eye movements and ASD, with the goal of designing specialized visual attention models. Baird *et al.* [2] introduced a dataset of vocalization recordings from children with ASD. They also evaluated classification approaches from the spectrogram of autistic speech instances. Their results suggest that automatic classification systems could be used as a tool for aiding in the diagnosis and monitoring of ASD in children.

Behavior analysis datasets. For the action recognition dataset of children with ASD, Pandey *et al.* [32] proposed a dataset of video recording actions and a technique to automate the response of video recording scenes for human action recognition. They evaluated their technique on two skill assessments with autism datasets and a real-world dataset of 37 children with ASD. Rehg *et al.* [34] introduced a publicly available dataset including over 160 sessions of child-adult interactions. They discussed the use of computer vision and machine learning techniques to analyze and understand children's social behavior in different contexts. They also identified technical challenges in analyzing various social behaviors, such as eye contact, smiling, and discrete behaviors. Rajagopalan *et al.* [33] explored the use of computer vision techniques to identify self-stimulatory behaviors in children with ASD. They also presented a self-stimulatory behavior dataset (SSBD) to assess the behaviors from video records of children with ASD in uncontrolled natural settings. Their dataset comprised 75 videos grouped into three categories: arm flapping, head banging, and spinning behaviors.

Comparison to ***MMASD****.* In Table 1, we compare our proposed dataset with related benchmarks. Overall, MMASD features diverse themes and scenes, capturing full-body movements with multi-modal features. In contrast, some works focused specifically on upper-body movements [10, 12, 31, 45]. MMASD also provides critical privacy-preserving features to represent body movements making it publicly accessible, while some works were conducted on raw videos that are either private or accessible only upon request [4, 10, 12, 31, 34, 35]. Additionally, it is collected from therapeutic interventions, reflecting participants' motor ability and providing valuable insights for treatment guidance.

3 METHOD

In the following sections, we describe the participants, procedure, and experimental settings of our proposed dataset. This study was approved by the University of Delaware's Institutional Review Board (IRB) # 637082 − 12.

3.1 Participants

We recruited 32 children with ASD (27 males and 5 females) from different races (Caucasian, African American, Asian, and Hispanic) through flyers posted online and onsite in local schools, services, and self-advocacy groups. Prior to enrollment, children were screened using the Social Communication Questionnaire [37], and their eligibility was determined by the Autism Diagnostic Observation Schedule-2 (ADOS-2) [28, 39] as well as clinical judgment. All the children were between 5 and 12 years old. Written parental consent was obtained before enrollment. The Vineland Adaptive Behavior Scales [36] were used to assess the children's adaptive functioning levels. In general, 82% of the participating children had delays in the Adaptive Behavior Composite. Specifically, 70% of them experienced communication delays, 80% had difficulties with daily living skills, and 82% had delays in socialization

3.2 Procedure

The study was conducted over ten weeks, with the pre-test and post-test being conducted during the first and last weeks of the study, respectively. Each training session was scheduled four times per week and lasted approximately 45 minutes. During the intervention, the trainer and adult model interacted with the child within a triadic context, with the adult model acting as the child's confederate and participating in all activities with the child. This triadic setting (child, trainer, and model) provided numerous opportunities for promoting social and fine motor skills such as eye contact, body gesturing and balancing, coordination, and interpersonal synchrony during joint action games.

All expert trainers and models involved were either physical therapists or physical therapy/kinesiology graduate students who had received significant pediatric training prior to their participation. The trainers and models were unknown to the children before the study. In addition to the expert training sessions, we also encouraged parents to provide two additional weekly sessions involving similar activities to promote practice. Parents were provided with essential instruction manuals, supplies, and in-person training beforehand. All training sessions were videotaped with the parents' consent and notification to the children, and the training diary was compiled by parents in collaboration with expert trainers. The general pipeline of training sessions had a standard procedure despite some unique activities across different themes. A welcoming and debriefing phase was present at the beginning and end of the data collection to help children warm up and get ready for the intervention, as well as to facilitate the subsequent data processing stage by providing time labels that indicate the segments to investigate.

Table 2: Statistics of our proposed MMASD dataset.

Description	Value
Number of data samples	1,315
Number of frames	244,679
Number of activity classes	11
Average video length (seconds)	7.0 ± 3.4
Average number of frames	186.1 ± 92.9
Resolution	$320 \times 240 \sim 1920 \times 1080$
FPS	$25 \sim 30$

3.3 Experiment Settings

All videos from triadic settings were recorded in a house environment with the camera pointed toward the participating child. Different tools were introduced to facilitate the training process depending on the theme of the intervention, for example, instruments and robots. Selected scenes in different themes of our proposed MMASD dataset are shown in Figure 2.

4 MMASD DATASET

MMASD includes 32 children diagnosed with autism of different levels. It covers three unique themes:

- Robot: children followed a robot and imitated bodily movements.
- Rhythm: children and therapists played musical instruments or sang together as a form of therapy.
- Yoga: children participated in yoga exercises led by therapists. These exercises included body stretching, twisting, balancing, and other activities.

Overall, MMASD comprises 1,315 video clips that have been meticulously gathered from intervention video recordings spanning more than 108 hours. It consists of 244,679 frames with an average duration of 7.15 seconds. The average data length in MMASD is 7.0 ± 3.4 seconds (186.1 ± 92.9 frames), with dimensions ranging from 320×240 to 1920×1080. Table 2 presents statistical information on MMASD. Depending on the conducted activity during the intervention, we further categorized all data into eleven activity classes as described in Table 3. Each activity class falls into a unique theme, as shown in Figure 2. MMASD also reports demographic and autism evaluation scores of all participating children, including date of birth, motor functioning score, and severity of autism.

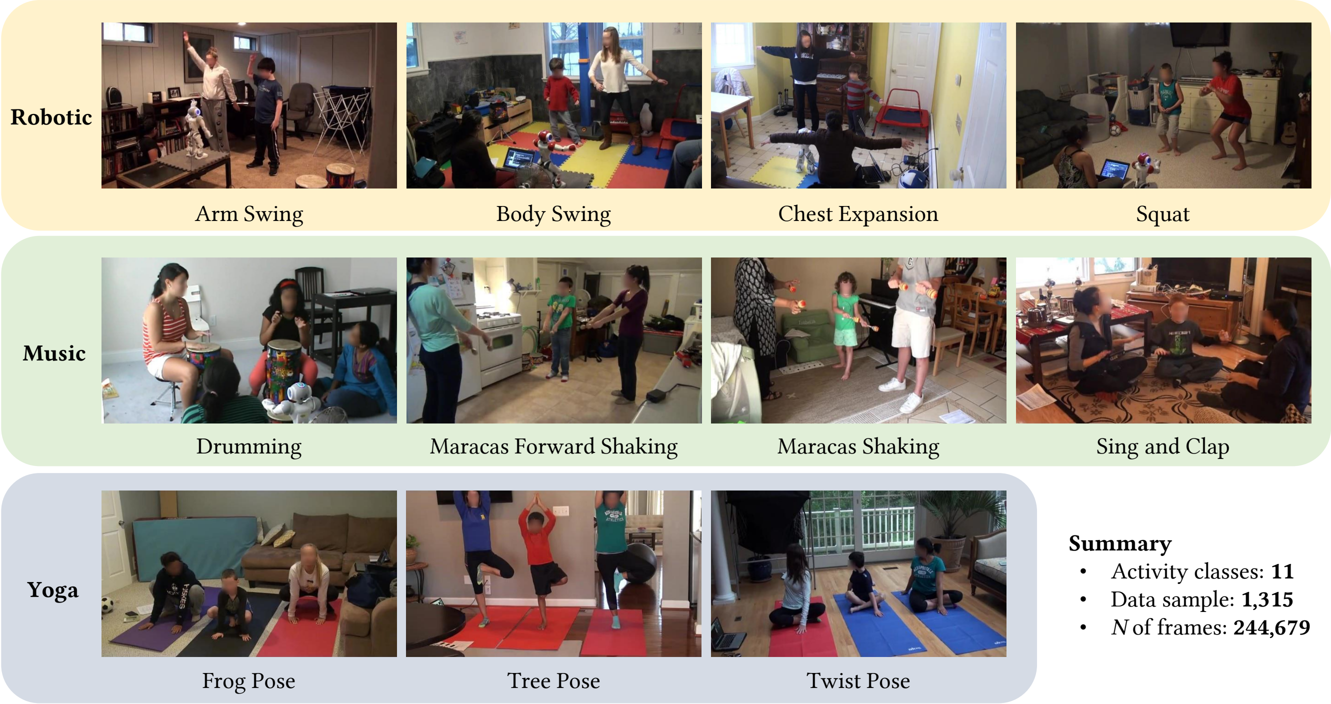

Figure 2: Sample scenes depicting various themes and activity classes present in the MMASD dataset.

Table 3: Description and distribution of all 11 activity classes in MMASD.

Activity Class	Activity Description	Count
Arm swing	The participant raises their left and right arm in succession while maintaining an upright posture.	105
Body swing	The participant swings their body left and right while stretching out both hands, one behind the other.	119
Chest expansion	The participant gradually opens and closes their chest.	114
Drumming	The participant plays either the snare or Tubano drum with one or both hands.	168
Frog pose	The participant widens their knees as far as possible, places their feet with the big toes touching their body, and assumes a shape like that of a frog in the kneeling position.	113
Maracas forward shaking	The participant shakes maracas back and forth, an instrument commonly appearing in Caribbean and Latin music.	103
Maracas shaking	The participant shakes maracas left and right in front of their chest.	130
Sing and clap	The participant sits on the ground while simultaneously singing and clapping, typically done at the start or end of an intervention.	113
Squat	The participant repeatedly performs a crouching stance with their knees bent.	101
Tree pose	The participant balances on one leg and places the sole of the other foot on the inner thigh, calf, or ankle of the standing leg in tree pose.	129
Twist pose	The participant sits with their legs crossed and twists their torso to one side, keeping their lower body stable and grounded.	120

4.1 Data processing

From original video recordings, we manually find out the start and end time stamps of a specific activity. Then we segmented the video into clips and categorized them by activity class. Clips shorter than three seconds were discarded. We also discarded noisy data due to video quality, lighting conditions, and body occlusion. Besides all the eleven activities in MMASD, there were some other activities with fewer examples. To ensure a balanced data distribution, we excluded all inadequate classes.

4.2 Data Annotation

We have four annotators that are well-trained in intervention understanding. The annotators had a comprehensive understanding of interventions and a cross-disciplined background in computer science and physical therapy. We exclusively assigned one activity class label to each video. Each annotator completed data annotation independently, and the final class label was determined by majority voting. However, original videos cannot be publicly shared due to privacy concerns. Therefore, we create data samples from every video clip by extracting selected features from the original scenes, including (1) optical flow, (2) 2D skeleton, and (3) 3D skeleton, respectively. All these features can maintain critical body movements while preserving privacy. Section 4.3 explained all selected features in detail.

In addition to the motion-related features mentioned above, we also reported clinician evaluation results such as motor functioning score, and severity of autism for each participating child. ADOS-2 is a standardized assessment tool used to evaluate individuals suspected of having ASD. It is used in conjunction with other diagnostic information to help clinicians determine whether an individual meets the criteria for an ASD diagnosis. ADOS-2 includes several modules, each designed for individuals of different ages and language abilities and includes a series of activities and tasks that are used to observe key features of ASD, such as social communication skills and repetitive behaviors. The ADOS-2 scores are based on an individual's performance during activities and tasks specific to their module and can range from 0 to 10 or higher depending on the algorithm used, with higher scores indicating more severe ASD symptoms. Moreover, we reported the ADOS comparison score, a continuous metric ranging from 1 to 10 that describes the severity of a child's autism symptoms compared to children with ASD of similar age and language levels [16]. Low comparison scores are indicative of minimal evidence of autism symptoms, whereas high scores are indicative of severe autism symptoms.

The contrast between the ADOS-2 score and the ADOS comparison score is worth mentioning. The ADOS-2 score reflects an individual's raw score on the ADOS-2 assessment tool, while the ADOS comparison score is a statistical measure that compares an individual's performance to others of the same age and language level. The motor functioning score refers to an assessment of an individual's motor skills and abilities and is evaluated based on the children's level of independence in daily living skills [37, 38]. It is on a scale of 1 to 3, while 1, 2, 3 represent low functioning (needing significant support), medium functioning (needing moderate support), and high functioning (needing less support), respectively. Finally, the severity of autism is determined by a comprehensive assessment that includes both the ADOS and motor function evaluation.

4.3 Multimodal Feature Extraction

In order to preserve critical details of movement while avoiding any infringement of privacy, we derived the subsequent features from the initial footage.

(1) Optical flow An optical flow is commonly referred to as the apparent motion of individual pixels between two consecutive frames on the image plane. Optical flow derived from raw videos (see Figure 1) can provide a concise description of both the region and velocity of a motion without exposing an individual's identity [6, 14, 15, 32].

(2) 2D Skeleton Skeleton data has an edge over RGB representations because it solely comprises the 2D positions of the human joints, which offer highly conceptual and context-independent data. This allows models to concentrate on the resilient aspects of body movements. 2D skeleton data has been widely applied to tasks relating to human behavior understanding, such as action recognition [43, 44], action quality assessment [26, 27] and beyond. An optimal way to acquire skeleton data is through the use of wearable devices and sensors that are affixed to the human body. However, in the context of autism research, it poses a substantial challenge as children may feel overwhelmed wearing these devices and experience anxiety. As a result, the skeleton data extraction process is carried out by pre-trained pose detectors based on deep neural networks.

(3) 3D Skeleton Similar to 2D skeleton data, 3D skeletons instead represent each key joint with a 3D coordination, introducing an additional depth dimension. Since all the data was collected using a single RGB camera, we also completed this process with the help of deep neural networks.

The technical details and tools utilized for feature extraction can be found in Section 4.5.

4.4 Data Format

Suppose the original video clip includes N participants (child, trainer, and assistant) is composed of L frames, and the height and width of each frame are H and W, respectively. As discussed above, each *data sample* consists of four distinct components, with data dimension demonstrated in braces:

- Optical flow $(L-1, H, W)$: saved as *npy* files [18].
- 2D skeleton $(L, N, 17, 2)$: 2D coordinates of 17 key joints, following COCO [9] format, saved as *JSON* files.
- 3D skeleton $(L, N, 24, 3)$: 3D coordinates of 24 key joints, following ROMP [40] format, saved as *npz* files.
- Demographic and clinical evaluation for ASD $(9,)$: including nine attributes such as participant ID, date of birth, chronological age, social affect score, restricted and repetitive behavior score, motor functioning score, and severity of autism, saved as *CSV* files.

4.5 Implementation Details

4.5.1 Optical flow: The Lucas-Kanade [29] method is used for our study. It is a popular technique used in computer vision to estimate the motion of objects between consecutive frames. The method assumes that the displacement of the image contents between two nearby instants is small and approximately constant within a neighborhood of the point under consideration. By solving the optical flow equation for all pixels within a window centered at the point, the method can estimate the motion of objects in the image sequence. Overall, the Lucas-Kanade optical flow method is an effective and preferred technique for estimating motion in various computer vision applications.

4.5.2 2D skeleton: The OpenPose method [5] is used to extract 2D skeletons from our dataset of human action videos. OpenPose is a powerful tool for body, face, and hand analysis, developed by Carnegie Mellon University, and is based on Convolutional Neural Networks (CNNs). It is a real-time multi-person key-point detection library that can accurately detect the key points of a human body, including joints and body parts, from an image or video feed. Initially, it predicts confidence maps for every body part and subsequently associates them with distinct individuals via Part Affinity Fields. The library is open-source and written in C++ with a Python API, which makes it easy to use and integrate into various computer vision applications.

4.5.3 3D skeleton: We utilized the Regression of Multiple 3D People (ROMP) proposed by Sun *et al.* [40], a state-of-the-art technique to estimate the depth and pose of an individual from a single 2D image. The authors proposed a deep learning-based approach that is based on a fully convolutional architecture, which takes an input image and directly predicts the 3D locations of the body joints of the person(s) present in the image. This is achieved by directly estimating multiple differentiable maps from the entire image, which includes a Body Center heatmap and a Mesh Parameter map. 3D body mesh parameter vectors of all individuals can be extracted from these maps using a simple parameter sampling process. These vectors are then fed into the SMPL body model to generate multi-person 3D meshes.

In our study, we employed the code and pre-trained model shared by the authors and used it on our dataset to suit our specific needs. By utilizing this method and applying it to our own data, we obtained 2D and 3D coordinates of key joints of the person(s). Since it is suitable for occluded scenes and noisy data, ROMP demonstrated its ability to successfully identify and represent the dynamics of our multi-class, multi-person activities.

5 DISCUSSION

This section delves into the challenges, insights, and future opportunities of MMASD dataset. As the experiments were conducted in real-world settings in children's homes, we faced common computer vision challenges, including varying video quality, illumination changes, cluttered backgrounds, and pose variations. Notably, in the feature extraction stage, we encountered pose detection failures in challenging scenarios, such as body occlusion and participants moving out of the scene. The intrinsic video quality limitation of MMASD also restricted us from capturing subtle and fine-grained features, such as facial expressions.

Moreover, it is imperative to conduct in-depth investigations into domain-specific challenges. For instance, in standard benchmarks for human activity recognition, typically developing individuals exhibit dominant and continuous actions with similar intensity. However, data on MMASD may not solely contain the target behavior throughout the entire duration, as impromptu actions or distractions may (and will) occur during therapy sessions for children with ASD. Furthermore, unlike prevailing benchmarks that collect ground truth skeleton data by attaching sensors to the human body, MMASD generates skeleton data by means of pre-trained deep neural networks. This is because children with autism have limited tolerance for external stimuli, and the presence of sensors on their bodies may cause them to become anxious, agitated, or exhibit challenging behaviors. Consequently, the skeleton data's reliability in MMASD depends on the performance of the underlying pose detectors. In addition, children with autism can exhibit varying motor functions, resulting in different intensity levels and completion rates for the same activity.

There are several directions for future work that are worth exploring. Firstly, further research can be conducted to develop and compare machine learning models on the MMASD dataset for various tasks, such as action quality assessment [24], interpersonal synchrony estimation [26, 27], and cognitive status tracking [8, 23]. This can help establish benchmark performance and identify state-of-the-art methods for analyzing the full-body movements of children during therapeutic interventions. In addition, new approaches can be investigated to overcome pose detection failures in MMASD. For example, by introducing pose uncertainty [27] or attention mechanism to assign higher weights to more reliable body joints. Furthermore, the MMASD dataset can be expanded in several aspects. This includes new features such as mutual gaze [17], or additional annotations including movement synchrony scores or task-specific clinician evaluations. Also, the MMASD dataset can be expanded by the inclusion of (age- and gender-matched) typically developing children data for developing more comprehensive and contrastive models. Finally, efforts can be made to dataset augmentation via existing benchmarks not limited to the autism domain by matching samples with similar motion features [32], which can significantly expand the scale of the autism dataset.

6 CONCLUSION

Autism research has been greatly facilitated by machine learning techniques, which offer cost-effective, non-invasive, and accurate ways to analyze various aspects of children's behavior and development. However, the lack of open-access datasets has posed challenges to conducting fair comparisons and promoting sound research practices in the field of autism research. In this paper, we have proposed the MMASD, a privacy-preserving publicly accessible multimodal ASD children dataset ($N = 32$). The dataset features diverse, hand-annotated clips from over $100hrs$ of raw videos from the play therapy interventions. Our dataset includes multimodal features such as 2D & 3D skeleton data, optical flow, demographic data, and clinical rating, offering a confidential data-sharing approach that can maintain critical full-body motion information. Moreover, each scene in our dataset depicts the same activity performed by a child and one or more therapists, providing a valuable template for comparing typically developing individuals with children with ASD. The open-access MMASD dataset distinguishes itself from existing works by utilizing privacy-preserving multimodal features to provide comprehensive representations of full-body movements across diverse therapeutic social activities.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude for our research team, therapists and students who conducted the therapy sessions (Sudha Srinivasan, Maninderjit Kaur, and Isabel Park). We also thank therapy participants and their caregivers for contributing to the MMASD dataset collection. This work is partly supported by our sponsors, Amazon Research Awards Program, University of Delaware AI Center of Excellence, the National Institute of General Medical Sciences (NIGMS, P20 GM103446E), the National Institutes of Mental Health (NIMH, 5R21MH089441-02, 4R33MH089441-03), and Autism Speaks (Grant #8137). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Jon Baio, Lisa Wiggins, Deborah L Christensen, Matthew J Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, Walter Zahorodny, Cordelia Robinson Rosenberg, Tiffany White, et al. 2018. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries* 67, 6 (2018), 1.

[2] Alice Baird, Shahin Amiriparian, Nicholas Cummins, Alyssa M Alcorn, Anton Batliner, Sergey Pugachevskiy, Michael Freitag, Maurice Gerczuk, and Björn Schuller. 2017. Automatic classification of autistic child vocalisations: A novel database and results. (2017).

[3] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*. 354–361.

[4] Erik Billing, Tony Belpaeme, Haibin Cai, Hoang-Long Cao, Anamaria Ciocan, Cristina Costescu, Daniel David, Robert Homewood, Daniel Hernandez Garcia, Pablo Gómez Esteban, et al. 2020. The DREAM Dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy. *PloS one* 15, 8 (2020), e0236939.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[7] Shi Chen and Qi Zhao. 2019. Attention-Based Autism Spectrum Disorder Screening With Privileged Modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[8] Vuthea Chheang, Rommy Marquez-Hernandez, Megha Patel, Danush Rajasekaran, Shayla Sharmin, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, and Roghayeh Leila Barmaki. 2023. Towards Anatomy Education with Generative AI-based Virtual Assistants in Immersive Virtual Reality Environments. *arXiv preprint arXiv:2306.17278* (2023).

[9] Moreno I Coco and Rick Dale. 2014. Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in psychology* 5 (2014), 510.

[10] Geraldine Dawson, Kathleen Campbell, Jordan Hashemi, Steven J Lippmann, Valerie Smith, Kimberly Carpenter, Helen Egger, Steven Espinosa, Saritha Vermeer, Jeffrey Baker, et al. 2018. Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder. *Scientific reports* 8, 1 (2018), 17008.

[11] Ryan Anthony J de Belen, Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. 2020. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry* 10, 1 (2020), 333.

[12] Marco Del Coco, Marco Leo, Pierluigi Carcagnì, Francesca Fama, Letteria Spadaro, Liliana Ruta, Giovanni Pioggia, and Cosimo Distante. 2017. Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot. *IEEE Transactions on Cognitive and Developmental Systems* 10, 4 (2017), 993–1004.

[13] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. 2019. A dataset of eye movements for the children with autism spectrum disorder. In *Proceedings of the 10th ACM Multimedia Systems Conference*. 255–260.

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1933–1941. https://doi.org/10.1109/CVPR.2016.213

[15] Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai. 2020. An asymmetric modeling for action assessment. In *European Conference on Computer Vision*. Springer, 222–238.

[16] Katherine Gotham, Andrew Pickles, and Catherine Lord. 2009. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of autism and developmental disorders* 39 (2009), 693–705.

[17] Zhang Guo, Vuthea Chheang, Jicheng Li, Kenneth E Barner, Anjana Bhat, and Roghayeh Barmaki. 2023. Social Visual Behavior Analytics for Autism Therapy of Children Based on Automated Mutual Gaze Detection. In *Proceedings of the International Conference on Cooperative and Human Aspects of Software Engineering* (Orlando, Florida) *(CHASE '23)*.

[18] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[19] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. 2011. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence* 34, 9 (2011), 1704–1716.

[20] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. 2015. Dense 3D face alignment from 2D videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–8.

[21] F De la Torre, WS Chu, X Xiong, F Vicente, X Ding, and J Cohn. 2015. Intraface. In *IEEE International Conference on Face and Gesture Recognition*.

[22] Ivan Laptev. 2005. On space-time interest points. *International journal of computer vision* 64 (2005), 107–123.

[23] Jicheng Li, Roghayeh Leila Barmaki, Li Zhu, Korosh Vatanparvar, Migyeong Gwak, Jilong Kuang, and Alex Gao. 2023. Advancements in Face Alignment Evaluation for Contact-less Vital Sign Detection. In *2023 IEEE-EMBS International Conference on Body Sensor Networks: Sensor and Systems for Digital Health (BSN)*.

[24] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. 2021. Improving the Movement Synchrony Estimation with Action Quality Assessment in Children Play Therapy. In *Proceedings of the International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. 397–406.

[25] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. 2021. A Two-stage Multi-modal Affect Analysis Framework for Children with Autism Spectrum Disorder. In *Proceedings of the AAAI-21 Workshop on Affective Content Analysis* (New York, USA). 1–8. http://ceur-ws.org/Vol-2897/AffconAAAI-21_paper1.pdf

[26] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. 2022. Dyadic Movement Synchrony Estimation Under Privacy-preserving Conditions. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 762–769.

[27] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. 2022. Pose Uncertainty Aware Movement Synchrony Estimation via Spatial-Temporal Graph Transformer. In *Proceedings of the International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. 73–82.

[28] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. 2000. The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders* 30, 3 (2000), 205–223.

[29] Bruce D Lucas and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vol. 2. 674–679.

[30] Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, and Cristian Sminchisescu. 2018. 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2158–2167.

[31] Katherine B Martin, Zakia Hammal, Gang Ren, Jeffrey F Cohn, Justine Cassell, Mitsunori Ogihara, Jennifer C Britton, Anibal Gutierrez, and Daniel S Messinger. 2018. Objective measurement of head movement differences in children with and without autism spectrum disorder. *Molecular autism* 9 (2018), 1–10.

[32] Prashant Pandey, Prathosh AP, Manu Kohli, and Josh Pritchard. 2020. Guided Weak Supervision for Action Recognition with Scarce Data to Assess Skills of Children with Autism. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (Apr. 2020), 463–470. https://doi.org/10.1609/aaai.v34i01.5383

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

[34] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. 2013. Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3414–3421.

[35] Giuseppe Riva, Eleonora Riva, et al. 2020. DE-ENIGMA: Multimodal Human-Robot Interaction for Teaching and Expanding Social Imagination in Autistic Children. *Cyberpsychology, behavior and social networking* 23, 11 (2020), 806–807.

[36] Sara S Sparrow and Domenic V Cicchetti. 1989. *The Vineland adaptive behavior scales.* Allyn & Bacon.

[37] Sudha M Srinivasan, Inge-Marie Eigsti, Timothy Gifford, and Anjana N Bhat. 2016. The effects of embodied rhythm and robotic interventions on the spontaneous and responsive verbal communication skills of children with Autism Spectrum Disorder (ASD): A further outcome of a pilot randomized controlled trial. *Research in autism spectrum disorders* 27 (2016), 73–87.

[38] Sudha M Srinivasan, Maninderjit Kaur, Isabel K Park, Timothy D Gifford, Kerry L Marsh, and Anjana N Bhat. 2015. The effects of rhythm and robotic interventions on the imitation/praxis, interpersonal synchrony, and motor performance of children with autism spectrum disorder (ASD): a pilot randomized controlled trial. *Autism research and treatment* 2015 (2015).

[39] Sudha M Srinivasan, Isabel K Park, Linda B Neelly, and Anjana N Bhat. 2015. A comparison of the effects of rhythm and robotic interventions on repetitive behaviors and affective states of children with Autism Spectrum Disorder (ASD). *Research in autism spectrum disorders* 18 (2015), 51–63.

[40] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. 2021. Monocular, One-stage, Regression of Multiple 3D People. In *ICCV*.

[41] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57 (2004), 137–154.

[42] Dennis Paul Wall, J Kosmicki, TF Deluca, E Harstad, and Vincent Alfred Fusaro. 2012. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry* 2, 4 (2012), e100–e100.

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence.*

[44] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 2021. 3D Human Pose Estimation with Spatial and Temporal Transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021).

[45] Andrea Zunino, Pietro Morerio, Andrea Cavallo, Caterina Ansuini, Jessica Podda, Francesca Battaglia, Edvige Veneselli, Cristina Becchio, and Vittorio Murino. 2018. Video gesture analysis for autism spectrum disorder detection. In *Proc. of International conference on pattern recognition (ICPR).* IEEE, 3421–3426.