Supplementary Materials: Loose Social-Interaction Recognition in Real-world Therapy Scenarios

Abid Ali[1 2] Rui Dai[1] Ashish Marisetty[1] Guillaume Astruc[1]
Monique Thonnat[1] Jean-Marc Odobez[3] Susanne Thümmler[1 2] Francois Bremond[1 2]

[1]INRIA [2]University Cote d'Azur [3]Idiap

Abstract

First, we discuss the child and clinician acquisition system for the Loose-Interaction dataset. Then we provide further details about the Loose-Interaction dataset and define each action class. Next we discuss experimental details. Finally, we explore the usefulness of automated diagnosis, and importance of loose-interaction in autism. Figure S1 provides an outline of the supplementary materials.

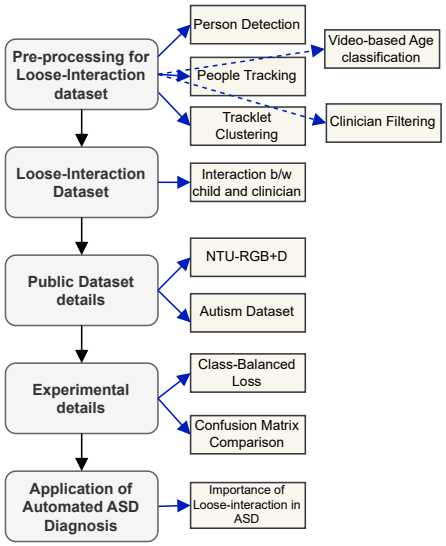

Figure S1. An outline of the supplementary materials.

1. Child and Clinician Acquisition for the Loose-Interaction

The goal is to encode the behavioural characteristics of two individuals in a complex environment. To achieve this, first, we acquire the child and clinician ("clinician", "psychologist", and "therapist" are used interchangeably) from the scene by applying an ensemble of SOTA methods including (People Detection, Tracking, Tracklet Clustering, Age-based Classification and Clinician Filtering). This part is crucial for Loose-Interaction dataset, specifically, where we have more than 2 people in a complex environment and we want to model only the child and clinician interactions. Figure S2 provides an overview of the whole system. In the Loose-Interaction dataset the role of clinician and child can be treated as a leader and an assistant, respectively.

1.1. Person Detection

The foremost step of our system is to detect all the people in the video. We adopt the **YOLOv5** [6] (a strong end-to-end object detection network) pre-trained on the **COCO** dataset [9]. The model takes an input $X_i^{h\times w\times c}$ and outputs the bounding box ($Y_i^{x,y,h,w}$), confidence score, and the class of the object (person in our case). The X_i represents the image that is being processed, while Y_i is the position of the bounding box of the person that is being detected in that image. We used an image of size 160×160 as the best choice to detect each person in our videos. We explored other detection algorithms including HOG [4], and YOLOv3 [11], but we found YOLOv5 the best among all.

1.2. People Tracking

The purpose of tracking is to provide each person's location with an ID within a video. We employ the **DeepSORT** [13] algorithm for this task. The DeepSORT algorithm is a combination of the Kalman filter for tracking the missing tracks and the Hungarian algorithm with a deep appearance descriptor to handle occlusion and viewpoint changes. The tracking algorithm takes the input of the detector in the form of $frame_i^{bbox_j}$, where $bbox_j$ are the bounding boxes in the i^{th} frame. The tracker provides an output of $frame_i^{tid_j}$ for each person's tracklet tid_j within that frame, j representing each person in the i^{th} frame. We compared DeepSORT with SORT [2] and FairMOT [14] to choose the best (deep-

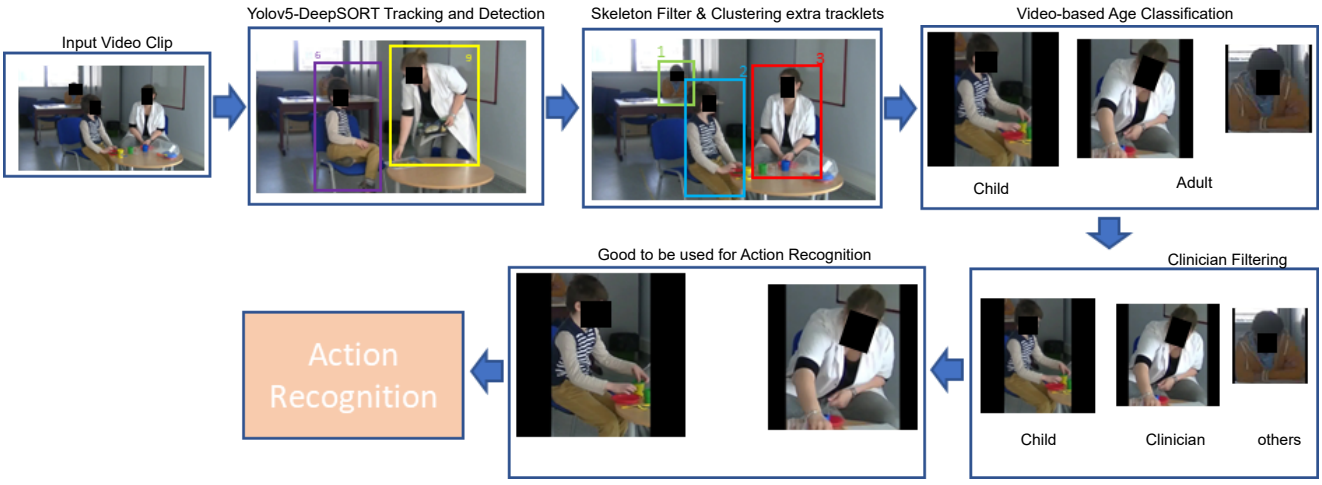

Figure S2. The block diagram of our entire system, from person detection to action classification.

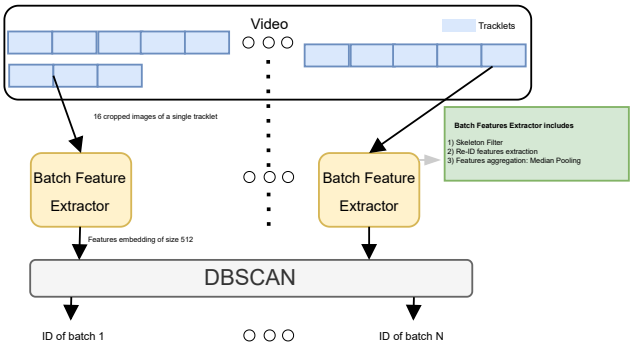

Figure S3. Overall Clustering approach.

SORT) among them for tracking.

1.3. Tracklet Clustering

In an ideal situation, a detector and a tracker are good enough to provide a single tracklet for each person in the video. However, this is not the case in real-world scenarios. Due to occlusions, camera movements, and spatial similarity (wearing similar clothes and/or parents' appearance similar to that of a child), it becomes very challenging to achieve a single tracklet ID for each person. Therefore, a clustering mechanism is introduced to cluster the IDs of a single person. This part not only merges multiple tracklets but also filters out unnecessary ones. We developed a skeleton-based filter using the **HRNet-w48** network [12]. Two different filters were applied to check the usability of the tracklet, one on the head and another on the top body (above the hips). The filter takes an input of $X_{b,h,w,3}$, where b, h, w represents the batch-size, height, and width, respectively. The Skeletal Filter checks if there are head joints (ears, nose, and eyes) or upper body joints (arms, and shoulders).

$$X_{b=16} = \begin{cases} \text{pass}, & \text{if } n\_sk \geq 8 \\ \text{discard}, & \text{otherwise} \end{cases} \quad (1)$$

Here, X_b represents a snippet of batch size 16 and n_sk is the number of skeletons. We only keep a skeleton if it has at least eight skeleton joints present as in Eq. 1.

Next, we fine-tuned **OSNet** [15] Re-Identification network on Loose-Interaction dataset with a triplet loss for re-id features extraction. Features of $X_{n\_sk,512}$ are taken from the last layer of OSNet [15] and are provided as input for the **DBSCAN** [8] clustering algorithm (as shown in Figure S3). A T-SNE projection of the clusters is given in Figure S4.

1.4. Video-based Age Classification

For the proposed two-stream architecture, it is necessary to identify the child (assistant) from the adults (parents and clinicians). Due to the resolution and movements of the camera of both the child and the clinician and the invisibility of the face, it was difficult to use an existing image-based age classification from the face images. To address these issues, similar to [1], we propose a video-based age classification network. Thanks to 3D-CNNs, we can model temporal information (hence, overcoming the issue of face occlusions). Also, analysing the entire body can help us capture important information about age (for example, shoulder width, body height, etc.). We design a custom **X3D** [5] architecture with an additional 512 MLP head and a BCE loss for age classification. The model inputs cropped images in the form $X^{b\times C\times h\times w\times t}$ to predict the classes (child vs. adult). Where b, h, w, c, t are batch size, width, height, channel, and temporal duration of the cropped tracklets, re-

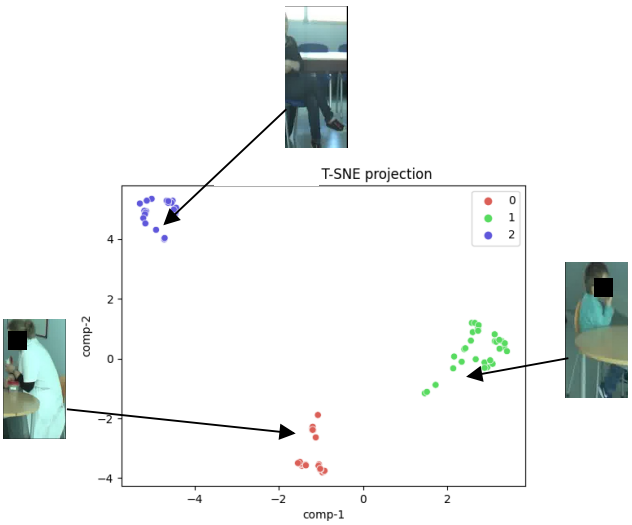

Figure S4. T-SNE projection of tracklet clusters for each person in the video.

spectively. We experiment with self-attention and different loss types to achieve the best results as shown in Table S1.

Network	Loss type	Precision/Recall	Acc.%
X3D*	BCE loss	0.92/0.92	92.3
X3D	Triplet loss	0.92/0.91	91.5
X3D	**BCE loss**	**0.93/0.93**	**93.5**

Table S1. Experimenting with different loss functions and architecture design. * represents X3D + self-attention

1.5. Clinician Filtering

We filter out unnecessary people (parents and other clinicians) from the videos using X3D [5]. We extract the CNN features for each video from the last CNN layer of the X3D [5] model in the form of $feats^{N\times c}$, where N represents the number of tracklets, and c is the features of the $conv_5$ layer. Taking log of the $L2$ norm for each tracklet within a video, we obtain the energy information per tracklet. The estimated L2-norm (energies) is sorted in descending order to pick the first ID (clinician) with the highest motion.

2. Loose-Interaction Dataset

The dataset is recorded while assessing children for possible ASD following the ADOS-2 protocol. After cleaning, we ended up with 845 clipped videos that have 9 action classes. In this paper, a total of 87 unique children's hour-long videos were used out of 132 videos. Figure S5 provides a few samples from the entire dataset. Each interactive action is explained below along with descriptions of clinician observation during each activity.

- **Anniversary**: The anniversary activity involves celebrating a birthday with a child (either for a doll or for the child). During this activity, several actions are carried out such as *cutting the cake, putting the cake on plates, pretending to eat the cake, picking up cups and putting them down on the table, giving cups to the child and taking them back, drinking from cups and clapping*. The activity is played with cups, plates and kinetic sand (used as cake) along with a doll.

  The clinician assess shared pleasure, social openings and reciprocity of the child.

- **Ball/balloon**: In this activity, the clinician plays with the child using a ball or a balloon. The clinician either fills the balloon with air or throws a ball at the child. The child usually reacts to the balloon or ball with joy and excitement by *jumping, clapping and chasing it*. However, sometimes when playing with a balloon, the child may show fear by *covering their ears with their hands*, especially if the balloon pops.

  Therapist observe child emotions, motor behaviour, and unusual movements.

- **Bubbles**: The clinician creates bubbles for the child to play with. The child may react to the bubbles by *popping them or walking towards them, and or jumping*. However, some children may not show any reaction to the bubbles and may only smile.

  Therapist evaluate the child's ability to initiate joint-attention, and shared pleasure. They also notice their sensory behaviour in this activity.

- **Construction**: The child and the clinician play together with wooden boxes. They perform actions such as *picking up the boxes and stacking them on top of each other* to construct a pyramid or other shape.

  Analyse eye-contact, and gestures of child.

- **Demonstration**: In this activity, the clinician asks the children about their morning routine and requests them to demonstrate actions such as *brushing their teeth, washing their faces and hands, and using soap*. Although these objects (soap, sink, and toothbrush) do not exist in the activity, the child and clinician pretend to have them.

  How the child represent familiar actions with gestures. How he/she demonstrate a sequence of actions. Does he/she mime imaginary object or uses his/her body to represent an object.

- **Describing an image**: The child and clinician interact using a piece of paper with pictures on it. The paper could be a larger picture or small cards. During the session, actions such as *pointing at images, talking about*

Figure S5. Samples from the Loose-Interaction dataset. We choose the best visible samples to allow the readers to easily understand the action types. Best viewed in colour.

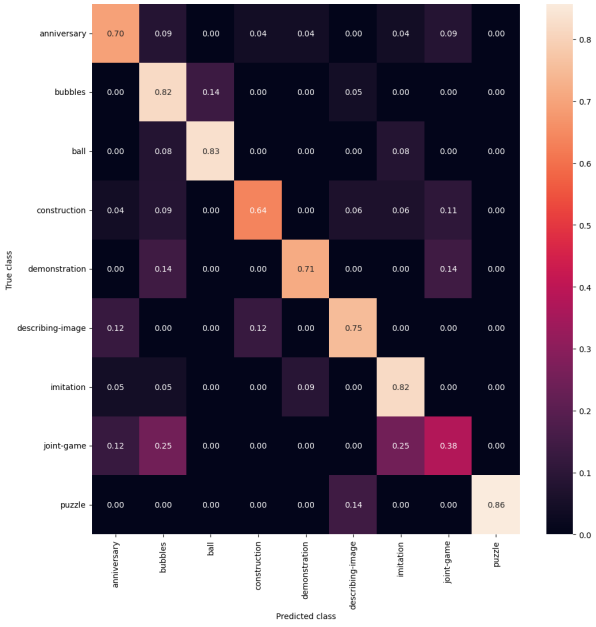

Figure S6. Visualisation of the confusion matrix of the proposed model on the Loose-Interaction dataset.

them, picking up small cards and putting them back are carried out.

Evaluate child's spontaneous language and communication as well as information about what interests him/her.

- **Imitation**: In this activity, the clinician first plays with an object in a specific way and then asks the child to imitate their gestures or actions. The clinician uses objects such as a toy frog, aeroplane and flower. Actions like *hopping the frog using hand gestures, flying an aeroplane in the air and smelling the flower are carried out*. Frog hopping is the most common activity performed by children.

  Assess child's social awareness, social skills, attention, and the use of miniature objects to imitate familiar actions.

- **Joint-game**: The child and the clinician play a game together. Some examples of games include *turning a toy together, playing with a dollhouse (with dolls sitting on toy sofas and tables), and playing with a toy car together*. The clinician and or the child can choose any game to play together.

evaluating the child interactive ability during play. how the child develop interactions and produce new initiatives beyond a direct response to clinician overtures. How the child propose new ideas or able to follow clinician's ideas.

- **Puzzle**: Playing a puzzle with the child interactively. The goal is to join small pieces to form a shape. During the activity, actions such as *picking up an object, putting it back, joining pieces together and exchanging objects between each other are carried out*.

  Analyse the child's facial expressions, creativity, and gaze.

2.1. Interaction between the child and clinician

Apart from the joint-game activity example, here we discuss a few more examples of loose interactions between the child and clinician. **Anniversary:** Their interaction is loose, because 1) the clinician (leader) is pouring drinks into cups, while the child (assistant) is cutting the cake or playing with a doll or clapping out of joy, 2) the clinician is assembling the plates and the child is drinking from the cup, etc. **Construction:** An interaction over creating a figure out of small boxes. They play the game together. Similar interaction can be observed in actions like imitation, describing-image and puzzle, etc. **Ball/Balloon:** Interaction can be observed in a catch-and-drop game, or in chasing the ball and pulling the ball/balloon away.

3. Public Dataset Details

3.1. Actions from NTU RGB+D Dataset

We combine interactive actions from the NTU RGB+D-60 and NTU RGB+D-120 datasets to create a tight interactive actions dataset. We used the 26 multi-action classes including A50 - A60 (NTU RGB+D 60), and A106 - A120 (NTU RGB+D 120). We use cross-subject settings for our experiments.

3.2. Autism Dataset

Autism dataset was proposed by [10] for ASD diagnosis in young children. The dataset has eight representative human action responses invoked through either listener response or imitation instructions. Specifically, action are chosen such as *move the table* and *arms up* for gross motor skill assessment, *lock hands* and *tap-ping* for fine motor skills, *rolly polly* for complex motor actions, *touch nose, touch head* and *touch ear* for identification of different body parts.

4. Experimental Details

4.1. More Ablations

We provide further ablations on loose-interactions dataset about backbone depth and number of attention heads. We noticed that the model accuracy decreases in using either 4 or 16 attention heads compared to 8 used in the model. The GLA module becomes too simpler for the complex interaction to learn when we use 4 attention heads. Similarly, using 16 attention heads leads to overfitting as the dataset is not large enough to learn such interactions as shown in Table S2. Furthermore, we also analyse the depth of backbone used. We choose 2 more backbones (X3D-S, and X3D-XL). With X3D-S as backbone we reach 70.4 % accuracy in Table S2. This validates a smaller backbone is enough to learn such interactions in a small dataset such as loose-interaction dataset. In contrast, with X3D-XL, the model performs worse. Initially, training the backbone with our method leads to overfitting. Therefore, we use a different training strategy by training the X3D-XL on the dataset first and then use the frozen backbone with our proposed modules which achieves 61.8% accuracy. This interesting analysis verifies our proposed modules are helping the backbone learn these interactions with trained together.

In addition, we further investigate the use of providing two different inputs, to analyse if we need two different inputs to the model as shown in Table S3. Using the same input to both streams increases the accuracy slightly compared to using a single stream.

Attention heads	Acc. (%)	Backbone depth	Acc. (%)
4	48.6	X3D-S	70.4
8	72.0	X3D-M	72.0
16	30.8	X3D-XL	61.8

Table S2. Analysis of using different attention heads in GLA module and backbone depth on loose-interaction dataset.

Same Inputs	Accuracy (%)
$X_{assistant}$	52.10
X_{leader}	58.64

Table S3. Analysis of using same input to both streams

4.2. Class-Balanced Loss

To combat the imbalance situation of our dataset, we adapted **Class-Balanced Loss** (CB Loss) [3]. CB loss introduces a weighting factor that is inversely proportional to the effective number of samples in each class. For an input

x with label $y \in \{1, 2, ..., C\}$, where C is the total number of classes, assume that the predicted output of the model for all classes is $z = \{z_1, z_2, ..., z_C\}$, where $z_i \in [0, 1]$ for all i, the loss function is denoted $\ell(z, y)$. For class i in n_i, the effective number of samples for that class is $E_{n_i} = (1 - \beta_i^{n_i})/(1 - \beta_i)$, where β is a weighting factor for the loss function, with hyper-parameter $\beta \in [0, 1]$ that controls how fast E_n grows as n increases. We use a combination of CB loss with focal loss, which can be written as:

$$\frac{1}{E_{n_y}} FocalLoss(z, y) = -\frac{1 - \beta}{1 - \beta_{n_y}} \sum_{i=1}^{C} (1 - p_i^t)^{\gamma} log(p_i^t) \tag{2}$$

$$CB(z, y) = \frac{1}{E_{n_y}} FocalLoss(z, y) \tag{3}$$

where n_y is the number of samples in the ground-truth class y, z is the predicted output and $p_i^t = sigmoid(z_i^t)$.

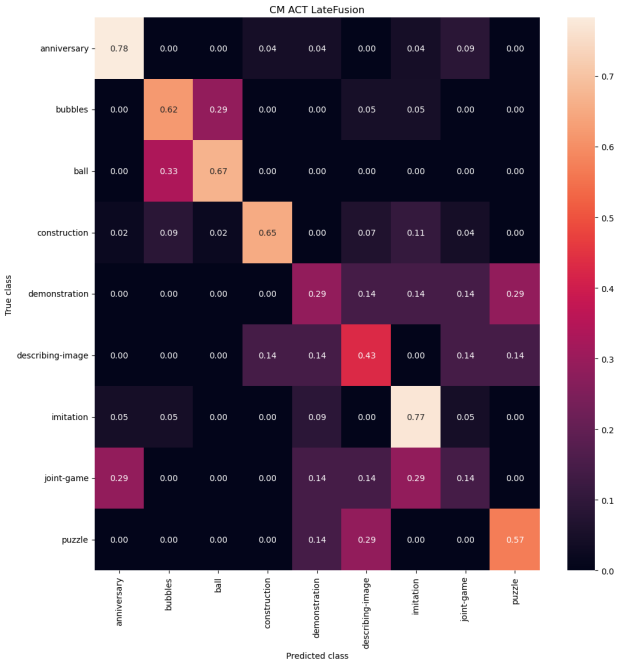

Figure S7. Confusion matrix of VideoMAE$_{fine_tuned}$ on the Loose-Interaction dataset.

4.3. Confusion Matrix Comparison

Our method achieves the highest accuracy (72.0) among all methods. We compare confusion matrix of our method (Figure S6) with VideoMAE$_{fine_tuned}$ as in Figure S7 for the Loose-Interaction dataset . Our method improves all action classes, especially *demonstration, describing-image, joint-game, and puzzle* with the highest margin compared to the late-fusion strategy for VideoMAE. Furthermore, we

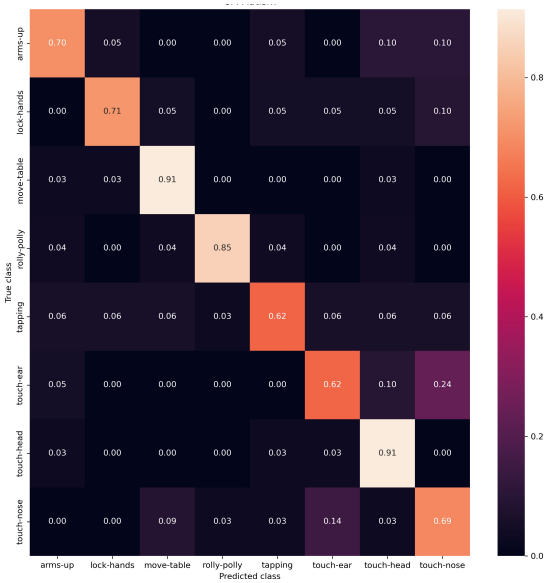

Figure S8. Visualisation of per-class accuracy of proposed network on the Autism dataset.

visualise the confusion matrix of our proposed method on the Autism [10] dataset as well. From the confusion matrix in Figure S8, we notice that some actions such as tapping, touch-ear, and touch-nose are highly dependent on the movements of the child only, and our model has a hard time recognising them.

5. Applications of Automated ASD Diagnosis

ASD diagnosis is a complicated challenge. The key factors are medical expertise, specialised diagnostic techniques based on interpreting child behaviour, parent interviews, long-term follow-ups, symptom inspection, and manual analysis. These evaluations take an extensive amount of time, and they demand laborious clinical procedures. Additionally, human assessments are subjective and inconsistent. Early studies suggested that abnormalities in social interactions, communication, and presenting repetitive behaviours could be the primary indicators of ASD [7].

An automated or semi-automated ASD diagnosis can be useful to countermeasure these issues. An automated ASD diagnosis tool can be useful in providing preliminary reports for clinicians for further diagnosis. Although, the automated system can be efficient one can not rely completely on such systems. These systems should be able to assist clinicians in their diagnosis. For instance, an automated system shall provide significant visual cues along with an autistic severity report for the clinician to make their final judgement.

Another application of automated ASD is to objectify the diagnosis. In other words, human assessments are subjec-

tive and can be compared with an AI-based assessment to make the final judgement.

On the contrary, such an AI ASD diagnosis system can be harmful if it is not experimented enough. For example, parents or clinicians relying solely on the AI diagnosis could lead to adverse results if the AI system is not efficient enough. Therefore, such systems should be tested extensively before making them available to the general public.

5.1. Importance of Loose Interaction in ASD

One of the major applications of this study is to assist clinicians in their diagnosis. Each ADOS-2 activity (anniversary, ball, construction etc.) is conducted to diagnose a specific autistic behaviour. For instance, an anniversary or joint-game is used to analyse the social interaction of the child. Similarly, ball/balloon and bubbles are used to trigger repetitive or stimming behaviour of an autistic kid. It is time-consuming and clinically requires arduous processes to look up these specific parts in a 2-hour long session. Our method can provide a summarised version of the session for the clinician to analyse. This summarising can help the clinician perform diagnosis efficiently.

Furthermore, in the future, we are interested in utilising this method for generating a social interaction report for child interactions. How involved are they in these social activities? One possible option can be using eye-gaze along with body gestures to generate a report for loose interactions.

References

[1] Abid Ali, Ashish Marisetty, and Francois Bremond. P-age: Pexels dataset for robust spatio-temporal apparent age classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8606–8615, 2024. 2

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 1

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[4] Navneet DARAL. Histograms of oriented gradients for human detection. *Proc. of CVPR, 2005*, pages 886–893, 2005. 1

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 2, 3

[6] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, (Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Nov. 2022. 1

[7] Leo Kanner et al. Autistic disturbances of affective contact. *Nervous child*, 2(3):217–250, 1943. 6

[8] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE, 2014. 2

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list. 1

[10] Prashant Pandey, Prathosh AP, Manu Kohli, and Josh Pritchard. 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):463–470, Apr. 2020. 5, 6

[11] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[13] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017. 1

[14] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 1

[15] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5056–5069, 2021. 2