Cognitive Computation (2025) 17:137
https://doi.org/10.1007/s12559-025-10481-7

RESEARCH

Multimodal Framework for Automatic Behavior Analysis of Children with Autism During ADOS-2

Bruno Carlos Dos Santos Melício[1] · Kaan Karaköse[1] · Ádám Fodor[1] · Linyun Xiang[1] · Viktor Varga[1] · Latha Soorya[2] · Emily Dillon[2] · Péter Kun[3] · András Sárkány[3] · Mohamed Chetouani[4] · Kristian Fenech[1] · András Lőrincz[1]

Received: 26 September 2024 / Accepted: 16 June 2025

Abstract
The rising prevalence of autism spectrum disorder, coupled with limited professional resources, highlights the urgency of developing efficient diagnostic tools. While standardized assessments exist, identifying subtle communication deficits, especially during multimodal interactions, remains time-consuming and prone to human error. To address this, we propose an automated behavior analysis framework that aims to support clinicians by accurately detecting both verbal and non-verbal communication markers. Specifically, we put forth a composite artificial intelligence framework that integrates various deep learning algorithms to analyze information from body and hand poses, object detection, tracking and manipulation, and speech. By combining these features with a rule-based system, we can identify events within the Autism Diagnostic Observation Schedule second edition, Construction Task, where participants initiate requests. These requests can be verbal, non-verbal or a combination of both resulting in multimodal interactions. Building on our prior work, this paper introduces a smart glass technology component, integrating gaze and blinking analysis, which are challenging for clinicians to monitor, given the multi-task nature of their role. These additions enable the detection of eye contact, a crucial social cue. Our approach allows us to recognize gestures, identify hand object manipulations, detect eye contact, and understand the natural language in clinician-participant interactions. We achieve 94% and 73% F-1 score, on verbal and non-verbal request detection, respectively, which may improve, as deep learning advances.

Keywords Autism spectrum disorder, Composite artificial intelligence · Behavior analysis · Blinking · Gaze · Eye contact

Introduction

Autism spectrum disorder (ASD) is a prevalent neurodevelopmental condition, affecting approximately 1 in 36 children in the USA, with a global impact on about 1% of the world's population [1]. Accurate diagnosis and assessment of ASD are crucial for timely intervention and support. However, identifying potential indicators or symptoms of ASD is inherently challenging, as it demands significant medical resources and can be time-consuming. One of the most complex aspects of ASD assessment lies in recognizing nuanced social communication behaviors, which often manifest through verbal and non-verbal cues. The difficulty in accurately identifying these behaviors highlights the need for automated, reliable, and interpretable tools to support clinicians, especially given the increasing prevalence of ASD and the scarcity of specialized resources.

The integration of artificial intelligence (AI) methods in behavior imaging [2–7] holds promise in automating the analysis of participant behavior, potentially enhancing efficiency and reducing the load on clinicians. However, due to the complexity and variability of ASD behaviors and the crucial need for interpretability, individual deep learning (DL) methods, often characterized by their black box nature, may not reliably provide accurate results. This limitation underscores the relevance of Composite AI (CAI). CAI, as defined by Gartner [8], refers to the fusion

✉ Kristian Fenech
fenech@inf.elte.hu

1 Department of Artificial Intelligence, Eötvös Loránd University, Budapest, Hungary

2 Department of Psychiatry, Rush University Medical Center, Chicago, IL, USA

3 Argus Cognitive, Inc, Hanover, NH, USA

4 Institute of Intelligent Systems and Robotics, CNRS UMR7222, Sorbonne University, Paris, France

Published online: 14 August 2025

of various AI techniques, both data-driven and rule based, to improve learning efficiency and broaden knowledge representations. By integrating various DL methods from different modalities with rules from a rule-based system (RBS), CAI addresses the interpretability challenge associated with DL models while leveraging domain-specific knowledge.

In this paper, we extend a CAI approach from our previous work [9] to detect verbal and non-verbal requests made by participants during semi-structured interactions, such as those observed in the Autism Diagnostic Observation Schedule second edition (ADOS-2). We introduce a novel component utilizing smart glass technology, incorporating gaze and blinking analysis to assess crucial social cues like eye contact, which play a significant role in communication.

We analyze video recordings from ADOS-2 sessions, recorded by both a wall-mounted camera and Tobii eye glasses worn by clinicians. ADOS-2 enables the assessment of autism levels across five modules, each designed for specific age groups and speech fluency levels. Our study focuses on Module 3, designed for verbally fluent children and young adolescents aged 5–14. This module was selected because effective analysis in this context requires participants with developed speech capabilities. Among the 14 tasks in Module 3, the Construction Task (CT) was chosen as it involves instances where participants initiate both verbal and non-verbal requests. A detailed description of the CT can be found on the "Dataset" section.

Detecting these requests is relevant in the context of autism diagnosis as effective communication is essential for social interaction, and difficulties in this domain are typical indicators of ASD. By analyzing how ASD individuals initiate and respond to requests, clinicians can gain insights into the subtle deficits in communication that characterize the disorder. This task proves to be challenging to analyze, due to the complexity of social interactions and the vast range of communication signs, which involve both verbal elements (spoken language) and non-verbal aspects (such as gestures, facial expressions, eye contact, and more).

To address these challenges, we propose a CAI framework, that operates in a modular fashion. It consists of the following three modules (see Fig. 1):

- Feature Extraction: Deep learning models from both vision and speech modalities are used to extract features from the environment, humans, and context.
- Episode Segmentation: uses rules to combine the features in order to identify and separate different events.
- Activity Interpretation: uses information from specific events of interest to classify different types of social cues.

This paper aims to demonstrate the capabilities of our proposed CAI framework by analyzing and interpreting complex social communication signs, both verbal and non-verbal, within the context of an ASD standardized test, ADOS-2. We emphasize the importance of gaze and blinking as fundamental features that may characterize both typically developing (TD) and autistic people during social interactions. In this context, detecting such cues, as well as interpreting situations where they may be absent or less noticeable, is crucial.

Thus, our CAI framework aims to provide the following benefits:

- Time efficiency: our modular framework automatically detects events of interest, allowing clinicians to focus on crucial moments rather than manually reviewing full recording sessions. This targeted analysis may significantly reduce the time required for ASD assessment.
- Supporting clinician decisions: Clinicians perform multitask roles, simultaneously observing, taking notes, and interacting with patients, which may cause them to overlook important signs. Our framework provides crucial information about subtle communication cues, such as

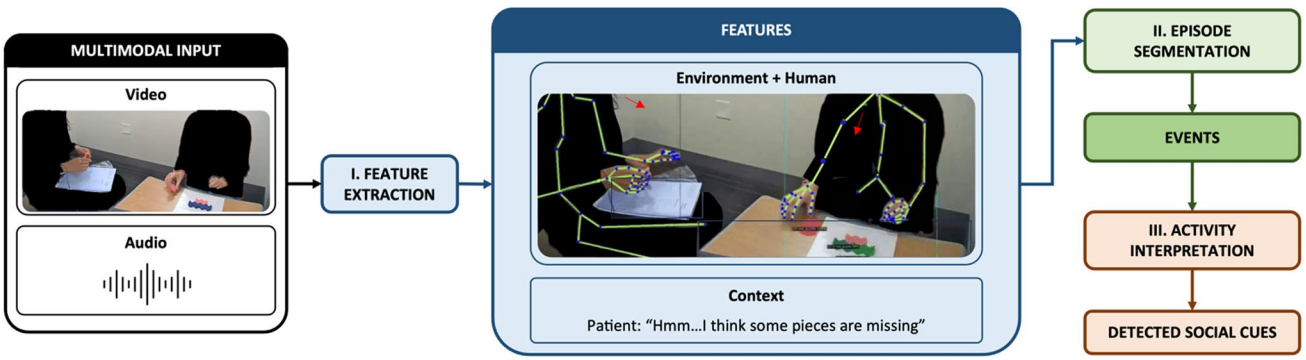

Fig. 1 An overview of our extended Composite AI framework v2. (I) Features about the environment, human, and context are extracted using different deep learning models; (II) features are combined using a set of rules from a rule-based system, to segment different events; and (III) another set of rules is used for detecting social cues within the events

gaze and blinking, which are challenging for humans to quantify. By offering comprehensive data, the system may aid clinicians in making more informed diagnostic decisions.

Related Work

Understanding and decoding the social behaviors of individuals with ASD present unique challenges due to their atypical social interaction, communication and behavior patterns. The severity and manifestation of autism symptoms vary significantly among individuals, resulting in diverse behaviors. A recent study [10] has underscored the complexity of social interactions and the crucial role of face-to-face interaction, highlighting the sensitivity of ASD individuals to social presence and its impact on their cognitive processes. Non-verbal cues like blinking and gaze are extensively researched due to their essential role in behavior analysis during social interactions, reflecting attention, focus, and engagement. Manual tracking of these cues is challenging, underscoring the need for automated methods.

AI methods [11–13] have shown promise in automating behavior analysis, accelerating the process, enhancing efficiency, and reducing clinician workload. Several studies [7, 14–17] utilize AI methods to detect verbal and non-verbal cues, aiming to understand the complexities of social interactions in individuals with autism. For instance, Celiktutan et al. [18] analyzed non-verbal autistic behavior during dyadic interactions using AI methods to extract individual and interpersonal features, while Chi et al. [19] focused on detecting distinct prosody patterns in speech recordings to automatically identify autism. Similarly, recent research [7] tracked voice, eye gaze, and facial expression features to classify individuals as autistic or not.

Atypical blinking patterns, such as elevated blink rates, have been linked with dopaminergic hyperactivity, providing insights into the neurobiological aspects of autism [20]. Additionally, deficits in temporal coordination of blinking with speech pauses have been observed in individuals with ASD [21]. AI methods for blinking analysis aim to quantify patterns of attentional engagement in autistic children based on facial orientation and blink rate. This study [22] reports that overall, autistic children had a higher mean blink rate compared to neurotypical children. Studies have shown that people with autism tend to avoid direct eye contact [23] and exhibit unconventional gaze patterns suggesting difficulties in selecting socially relevant information for attention [24]. AI methods, such as LAEO-Net [25] and Gaze360 [26], have been employed for estimating gaze and eye contact dynamics, enabling differentiation of distinctive sub-groups based on gaze behavior.

Our work applies deep learning models to extract vision and speech-based features, which may contain relevant information for characterizing both ASD and TD participants during social interactions. However, we refrain from classifying or diagnosing participants, focusing instead on detecting verbal and non-verbal cues to provide valuable additional information for clinicians. Furthermore, unlike existing approaches that analyze different features individually, our method integrates cues, such as blinking and gaze, into a Composite AI framework. This framework combines these features into modules along with rules, facilitating the tracking of the decision-making process.

Methods

Dataset

To assess our CAI framework, a proprietary dataset from RUSH Medical School, containing recordings of ADOS-2 sessions in English, was used. Informed consent and ethics approval were obtained through the Rush University Medical Center IRB. We focus on the Construction Task (CT) within Module 3 for children and adolescents with fluent speech. During the CT, clinicians guide participants through assembling puzzle pieces within a given shape; clinicians deliberately withhold certain pieces, prompting the participants to request additional ones. Participants' requests, whether verbal or non-verbal, play a crucial role in behavior assessment and may influence the evaluation of the task. The sessions are recorded by two sources: a wall-mounted camera and Tobii eye glasses worn by clinicians. The glasses provide a direct view of the participant and eye tracking data of the clinician. Each CT lasts approximately 2–3 min. The dataset comprises 29 participants (see details on Table 1), with some sessions discarded due to technical issues, resulting in 27 (20 ASD, 7 TD) usable Tobii videos.

Each session from the wall camera was manually annotated by two students involved in the research, with uncertain cases validated by two expert clinicians. Annotations include the start and end times of participant requests, along with the request type: verbal communication (spoken language),

Table 1 Details of participants, including sample size, gender, and age information. ASD, autism spectrum disorder; TD, typically developing

Group	Number of participants	Number of males	Number of females	Age mean (standard deviation)
ASD	22	15	7	9.7 (±1.8)
TD	7	4	3	8.6 (±1.7)
Total	29	19	10	9.4 (±1.8)

non-verbal gestures (hand or arm gestures), grabbing (taking puzzle pieces without a request), clinician (clinician provides puzzle pieces without participant request), or a combination (i.e., requests:{"start":["00:01:20", "00:02:51"], "end":["00:01:26", "00:03:00"], "type":["verbal", "gesture"]}). On average, each video session involves two requests (see Table 2).

Composite Artificial Intelligence

Composite artificial intelligence (CAI) is an emerging paradigm that combines data-driven approaches such as deep learning and traditional ones such as rule-based systems. By combining the strengths of both, CAI enhances model interpretability and efficiency, making it particularly effective for complex tasks like autism assessment, where domain-specific knowledge and transparency are critical. This hybrid approach provides interpretability, allowing one to track the decision-making process of the system, which is essential for building trust in AI, especially in fields like medicine. Our CAI framework consists of three modules: Feature Extraction, Episode Segmentation, and Activity Interpretation, each contributing to its robustness and effectiveness in detecting social cues.

Feature Extraction

This module uses deep learning models from both vision and speech modalities to capture diverse features that represent the environment, characterize the humans, and provide context. Most models are pre-trained with frozen weights, as they already perform well in our scenarios and due to the lack of ground truth for tasks such as body pose, objects detection and tracking, depth estimation, and more.

However, the hand pose and image segmentation models performed sub-optimally, leading us to fine-tune these models with our dataset to enhance their effectiveness for these specific tasks. Figure 2 shows all the models used.

Previous Features from Composite AI framework v1

In our initial work [9], we initiated the feature extraction process by detecting and tracking people and objects, essential for the Construction Task execution. Initially, YOLOv7 [27] was used for object detection. However, in this extended version, we transitioned to YOLOv8X [28] for improved precision, with an image size of $640 \times 640 \times 3$. The detection confidence thresholds were set at 0.9 for class people and 0.25 for the table and book classes. Non-max suppression was used with an IoU threshold of 0.45. For tracking, we use ByteTrack [29], which provides a unique ID for each object instance. We utilized the *bytetrack_x_crowdhuman_mot20* model weights and set a bounding box match threshold of 0.8.

Bounding boxes though useful for spatial information are insufficient for understanding behavior. Therefore, we leverage 3D body pose estimation methods, which provide keypoints of different body parts. We used a combination of MeTRABs [30] and MediaPipe [31] methods, which contains visibility scores, useful for overcoming occlusions. For MeTRAbs we make use of the largest backbone EfficientNetV2-L which outputs 24 joint keypoints in 2D and 3D. For MediaPipe [32] we use the *model_complexity* = 2 for the most accurate model but with slower inference. For the detection and tracking confidence, with use the default values of 0.5.

While the aforementioned body pose methods do provide hand keypoints, those are noisy and are significantly impacted by motion blur caused by the moving hands. Due to the significance of hands for expressing social cues such as hand gestures and their importance within the Construction Task for manipulating the puzzle pieces, the following method was developed using the steps below:

1. Hand segmentation: Fine-tuning Mask R-CNN [33] to segment hands
2. Body keypoints: Using Mediapipe keypoints to filter out segmentation errors (eliminates non-hand instances)
3. Tracking: using a modified version of MiVOS [34] for robust tracking, addressing occlusion issues
4. Hand keypoints: Use the segmented and tracked hands from the previous steps to enhance MediaPipe hand pose estimation [35]

For Mask R-CNN [33], we retain only predicted masks with a confidence greater than the threshold of 0.9. In the

Table 2 Number of requests by type per participant group. Comb.1: Verbal + Gesture combination; Comb.2: Verbal + Grab combination

Group	Verbal	Gesture	Grab	Comb. 1	Comb. 2	Clinician[1]	Total
ASD	30	1	7	3	1	6	48
TD	6	0	2	3	1	2	14
Total	36	1	9	6	2	8	62

[1] "Clinician" type indicates instances where the participant did not make any request. The clinician either attempted to prompt a request by asking if more pieces were needed or directly provided additional pieces. Such instances were excluded from the participant's evaluation

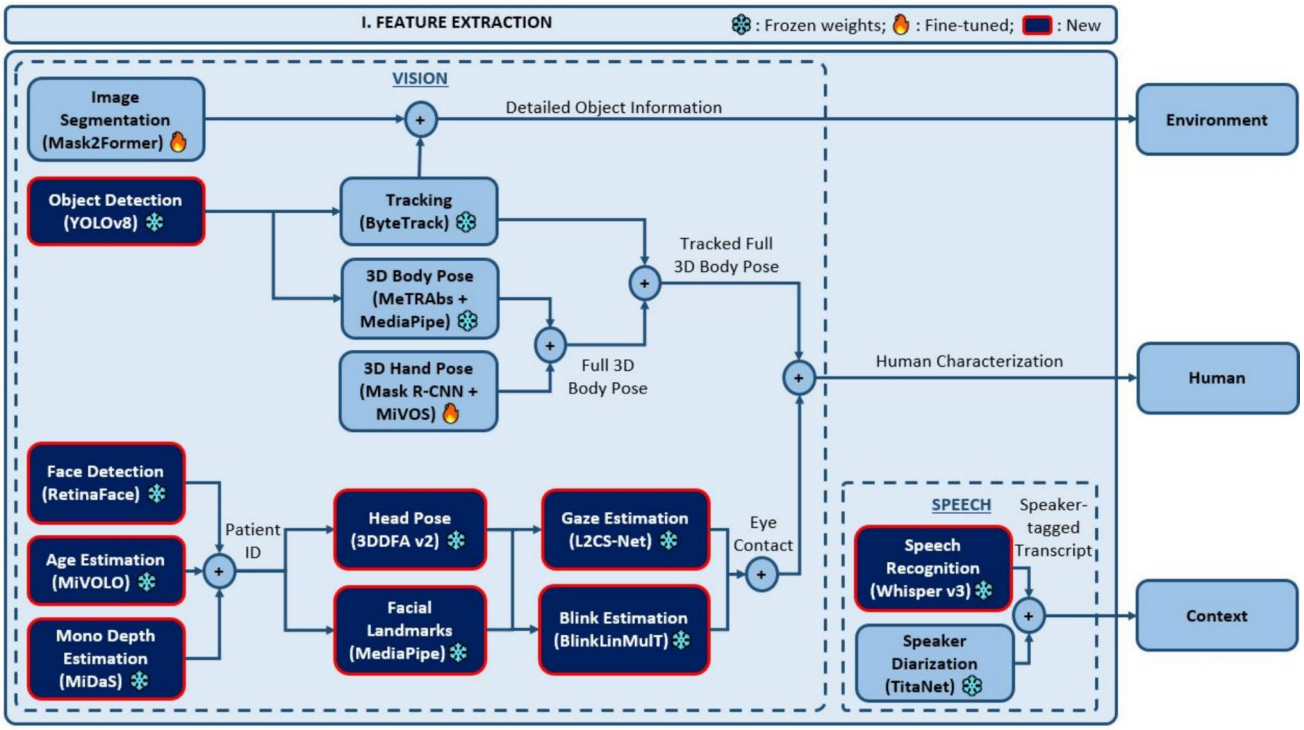

Fig. 2 Feature extraction module uses vision and speech models to extract features about the environment, human, and context. Models with frozen weights were trained on publicly available datasets, while fine-tuned models were further trained on our ADOS-2 dataset. In this extended version, additional models denoted by dark blue boxes with red borders are introduced

body pose component, we keep joints with a visibility score higher than 0.7. A distance threshold of 40 pixels between the center of predicted masks and the center of hand keypoints must be met to validate the masks. Tracking is then applied with a propagation value of 6 frames backwards. Finally, to obtain the 3D hand pose from MediaPipe Hands [35], we keep keypoints with a visibility score of at least 0.5.

To understand hand-object manipulations, precise information about the object is needed. Bounding boxes might include unwanted regions, therefore, segmentation was used to obtain the exact object size. Mask2Former [36] was fine-tuned for the puzzle piece object of the Construction Task, resulting in a robust method that can segment objects even if partially occluded by the hands. We used 13 videos (around 40,000 frames) where the puzzle pieces were manually annotated. We divided the data into 80% for training, 15% for validation, and 5% for testing. We used the ResNet101 [37] backbone, a batch size of 2, a maximum iteration of 5999, and a learning rate of 0.000025. The model was trained on 2×Titan RTX GPUs for 2 h. The total loss function combines classification loss and mask loss, as shown in Eq. (1). The λ_{cls} values are set to 2.0 according to [36].

$$L = L_{mask} + \lambda_{cls} L_{cls} \tag{1}$$

The classification loss is from [36] and the mask loss is calculated by adding binary cross-entropy loss and dice loss, as shown in Eq. (2), with both λ_{ce} and.

λ_{dice} set to 5.0 according to [36].

$$L = \lambda_{ce} L_{ce} + \lambda_{dice} L_{dice} \tag{2}$$

Finally, speech and dialogues, integral to social interactions, were also incorporated into our previous version of the framework. Initially, we used Whisper v2 [38] for speech recognition, but now we transitioned to Whisper v3 due to its smaller word error rate. However, as we need to identify the participant's request, we used TitaNet-L [39] for speaker diarization, in order to distinguish between speakers.

For more details about the v1 features, please refer to our prior work [9]. New added features are described in the next subsections.

Participant Tracking via *Face Detection, Age Estimation, and Depth Estimation*

In the ADOS-2 test, a clinician assesses a participant, often accompanied by a parent. To analyze the participant's behavior in recorded sessions, it is essential to accurately identify them amidst others.

Our approach employs a multi-step process combining three methods for effective participant tracking throughout the session. Initially, we utilize the RetinaFace [40] face detector to extract face bounding boxes from each frame. Given that Module 3 typically involves the youngest individual, we leverage MiVOLO [41] with default parameters to estimate the age of each detected face. By consistently selecting the youngest person over time, we achieve reliable participant tracking.

However, the clinical setting often includes mirrors, leading to duplicate faces. To address this, we integrate depth estimation using the MiDaS model [42]. Faces beyond a specified distance threshold are excluded from tracking. The threshold of 25 was determined empirically through trial and error, considering that the depth associated with mirror reflections typically falls within the range of 0 to 30 in the gray-scale depth map.

This combination of methods proves to be a robust solution for seamlessly tracking the participant throughout the session, even in the presence of varying ages and reflective surfaces. This demonstrates the effectiveness of our approach in real-world scenarios.

Blink Detection

To enhance the feature extraction capabilities of the CAI framework, we integrated BlinkLinMulT [43] for eye blink detection, leveraging both low- and high-level features from video sequences through linear complexity attention mechanisms.

For precise eye landmark prediction, we utilized MediaPipe's Face Mesh [44], a residual neural network adept at predicting dense coordinates even with partial occlusion of the face. These landmarks facilitated the cropping of eye patches from face images detected by RetinaFace.

Subsequently, the eye aspect ratio (EAR) was computed using MediaPipe Iris [45], which predicts the pupil center and 4 points of the outer iris circle in 2D. Additional high-level features were derived from iris diameters and distances between the pupil and eyelid landmarks.

To capture head pose information, we employed 3DDFAv2 [46], providing yaw, pitch, and roll angles describing the head's orientation. These features enabled BlinkLinMulT to effectively address challenges posed by diverse head poses and lighting conditions. We modified some parameters of BlinkLinMulT to better fit our ADOS-2 dataset. The initial step involves head pose estimation, where the predicted yaw angle guides the eye state recognition process. Extreme yaw head poses may result in self-occlusion, making one eye invisible due to the nose. To mitigate this, if the predicted yaw angle exceeds a defined threshold (set to 25° in our experiments), only the visible eye is utilized. Conversely, if the yaw angle falls below the negative of the threshold, only the opposite eye is considered. When the yaw angle falls within this threshold range, predictions from both eyes contribute to the analysis.

BlinkLinMulT processes 0.5-s video clips by extracting features from non-overlapping moving windows. Extracted features are fed into the model to generate frame-wise estimation logits. Mean logits are calculated to consolidate information from both eyes. A sigmoid function is then applied to these logits, and the results are thresholded to determine binary eye states (open or closed). We set the eye state threshold to 0.8 for our dataset, found to be most reliable for Tobii video sessions. Consecutive frame-wise binary eye state estimates are merged to identify blink events.

Gaze Estimation and Eye Contact

Gaze estimation is conducted using video recordings from Tobii glasses worn by clinicians. This setup offers an optimal view of the participant's face, facilitating the determination of whether the participant is looking at the clinician. Given that the clinicians wear Tobii glasses, equipped with a camera, on their face, we determine if the participants are looking at the clinicians by assessing whether they are looking at the center of the camera lens.

In our extended framework, we utilize the L2CS-Net method [47] to obtain the gaze pitch and yaw angles for precise gaze estimation of each participant. Initially, the face bounding box of the tracked participant, along with the face landmarks and head pose obtained from the methods described in the "Methods" section, is employed as preprocessing steps.

For gaze assessment, default hyperparameters of L2CS-Net are utilized to derive the gaze vector, represented by yaw and pitch angles within a spherical coordinate system. These angles are then transformed into pixel space to determine gaze orientation toward the clinician. The norm of the gaze vector, representing gaze direction, is analyzed using a threshold of 0.45 to ascertain if the participant is directing attention toward the camera lens.

Gaze vector estimation is only required for the participant, as Tobii glasses provide gaze information about the clinician. Equipped with infrared light sources near the eyes, these glasses reflect light off the cornea and other eye components, captured by eye-tracking cameras. As the wearer shifts their eyes and focuses on various objects or points of interest, the eye-tracking system consistently records data, providing real-time information on the person's gaze coordinates for every frame. In this case, to determine if the clinician is looking at the participant, we estimate if the gaze point (x, y) pixel values are within the bounding box of the participant's face.

Finally, employing this method, mutual eye contact between the clinician and participant is identified when

both parties are observed to be looking at each other, based on the described methods above.

Combined Features: Synchronization

Each session in the dataset comprises two sources of unsynchronized videos: one recorded by a wall-mounted camera and another by a camera from the Tobii eye glasses worn by the clinician. While annotations regarding the participants' verbal and non-verbal requests were available for the wall camera, synchronizing the corresponding timestamps of requests in the Tobii camera recordings necessitated temporal alignment. We employ a dual synchronization approach comprising text-based and audio-based methods.

The text-based synchronization attempts to match text transcripts from the two sources, using the Whisper large-v3 [38] automatic speech recognition (ASR) model. Whisper transcribes the audio of each camera source and provides sentence-level timestamps (i.e., when the sentence started being spoken and when it ends). However, due to differences in the audio quality of the two sources, Whisper may transcribe the same sentence differently due to noise. Therefore, in order to match the transcripts from both camera sources accurately, word-level timestamps were calculated by using the Montreal Forced Alignment software [48]. With that, a word-wise match could be made. Specifically, we try to match the longest sequence of consecutive words transcribed in both sources.

While this method suffices for most recorded sessions, significant noise in the audio signal from the Tobii camera source occasionally leads to delays of up to 10 s in synchronizing the two camera sources, which is suboptimal for request detection. To enhance temporal accuracy, we refine transcript-based timestamps using a secondary technique based on audio features. We conduct time-delay analysis of the two audio signals by computing cross-correlation of the normalized audio sources with varying offsets.

Due to high levels of noise, the audio-based technique on its own often fails to accurately determine the time delay between the two signals. However, by limiting the calculation to a narrow range of possible offsets around the prediction of the transcript-based method, we could successfully refine the time delay coming from the prediction of the transcript-based method.

The combination of the two methods worked for most videos, with the exception of three with extreme noise, which were manually matched. Having the camera sources synchronized, the features extracted from each source can be combined and analyzed together (see Fig. 3).

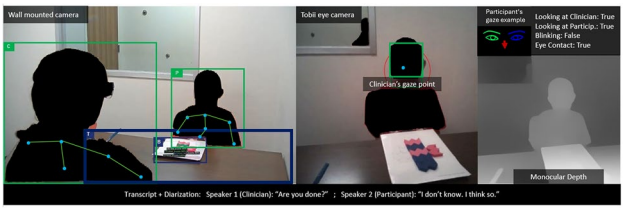

Fig. 3 Combined vision and speech features overlaid on images from two synchronized cameras. On the left, body pose, object detection, tracking, and segmentation are extracted from the wall-mounted camera. On the right, the clinician's gaze (blue dot), participant's gaze (red arrow), blinking, eye contact, and depth are extracted from the Tobii camera. On the bottom, transcript and diarization from speech are shown

Episode Segmentation

Episode segmentation identifies distinct events within the input, combining visual and speech cues with a rule-based system. It defines event intervals, including start and end points. We define four episodic components (see Fig. 4):

- Start: Sets the beginning of the task when people are present. Using the object features, it is denoted by the moment when one or more puzzle pieces are first placed on the table.
- Interaction: Identifies when people interact with each other or with objects. Relies on both vision and audio features such as hand pose, object detection/segmentation and speech. Using the hand and object features, the interaction is identified when at least one hand and one puzzle piece move together within the paper's region. This means that the participant has started to interact with the puzzle. This episode also takes into consideration the conversation that leads to the interaction, i.e., usually the participant starts engaging with the puzzle after the clinician has finished the verbal explanation of the task and/or has given explicit permission to start.
- Transition: The most important episode is the transition/request event, which denotes a moment where a social request is likely to happen. To identify it, we use a combination of the different extracted features. Specifically, we track the number of puzzle pieces, introduction of new speech topics in the conversation, shifts in gaze focus on persons or objects, and changes in body posture. We assume that these changes often lead to a request; thus, we mark these transitions as key episodes to analyze.
- End: Indicates the end of an interaction. Task-specific rules can be added to help determine when an interaction has ended. For our case, the end is identified

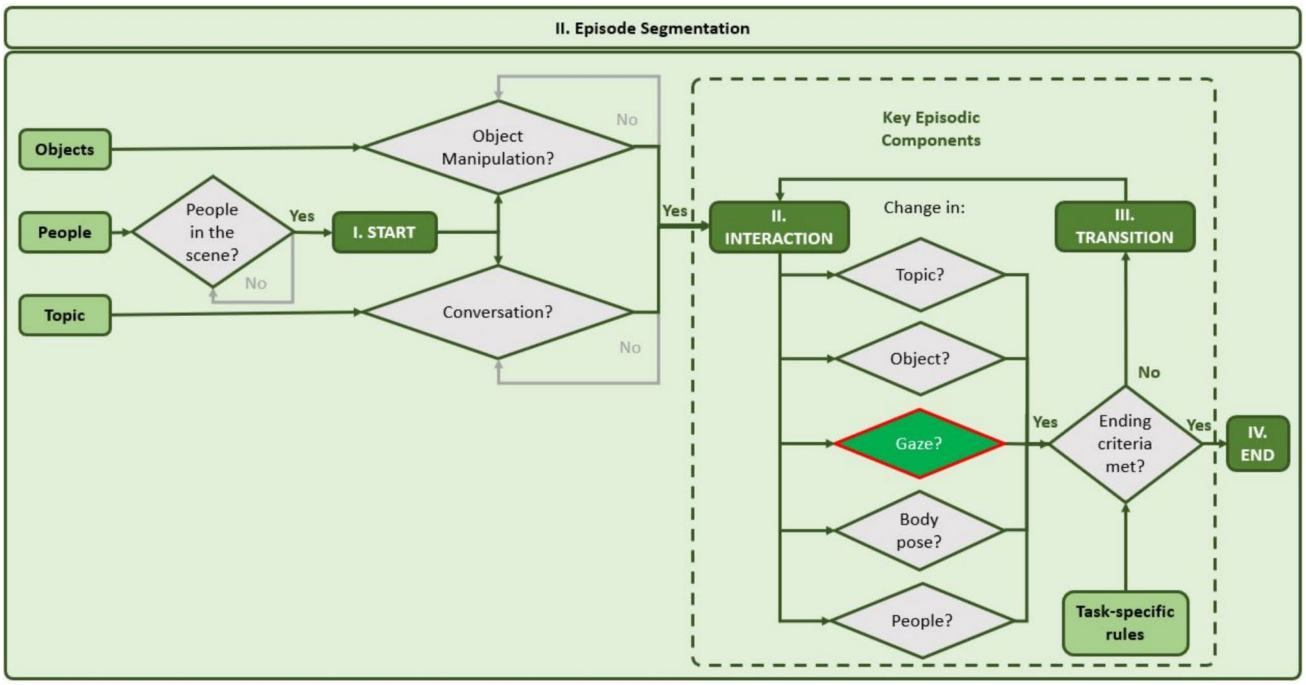

Fig. 4 Episode segmentation: given the previously extracted features about the objects, people, and topic, rules specific to the task are used to combine the features and detect events. Four generalized events (start of the task, interaction, transition, and end of task) are defined. During interaction events, participants engage either physically with objects or verbally with people. Any observable change during these interactions is saved as a key episode. We introduce gaze as one of the new features that may show important changes. The task concludes when the specified ending criteria are met

Table 3 Summary statistics of episode durations

Episode type	Total occurrences	Avg. number of episodes per video	Avg. duration (s)	Std. dev. of duration (s)
Start	27	1.0	7.4	11.8
Interaction	74	2.7	18.4	17.6
Transition	47	1.7	5.5	3.9
End	27	1.0	48.3	23.5

once all the puzzle pieces are detected on top of the paper and are not moving, or the clinician's hands are removing the pieces from the table.

The most relevant events are likely to occur between the interaction and transition episodes because something must trigger a change. Therefore, we constrain the detection of social cues (verbal and non-verbal) to these key moments, enabling us to search for specific events within a limited time frame and reduce false detections. Detailed information about each episode can be found in Table 3. Additionally, Fig. 5 illustrates the distribution of episode durations.

Activity Interpretation

With the analysis constrained to the "Transition/Request" episode, we combine features using classifiers and rules to detect verbal and non-verbal cues by recognizing gestures, identifying different hand-object manipulations, estimating eye contact and understanding natural language (see Fig. 6). This process, termed "activity interpretation," goes beyond just detection. The integration of rules to govern the features extracted from deep learning methods facilitates traceability of reasoning processes, enables a clearer understanding of decision-making mechanisms, and fosters human trust in the AI's outcomes. Clinicians can control the rules by adding domain-specific knowledge.

Natural Language Understanding

In social interactions, humans converse with one another and may need something from the other, which could be the main purpose of the interaction. Thus, understanding the person's intention is crucial for effective communication. The previously extracted transcripts and speaker diarization, are utilized to identify speakers and their utterances. We employ a vocabulary of common request

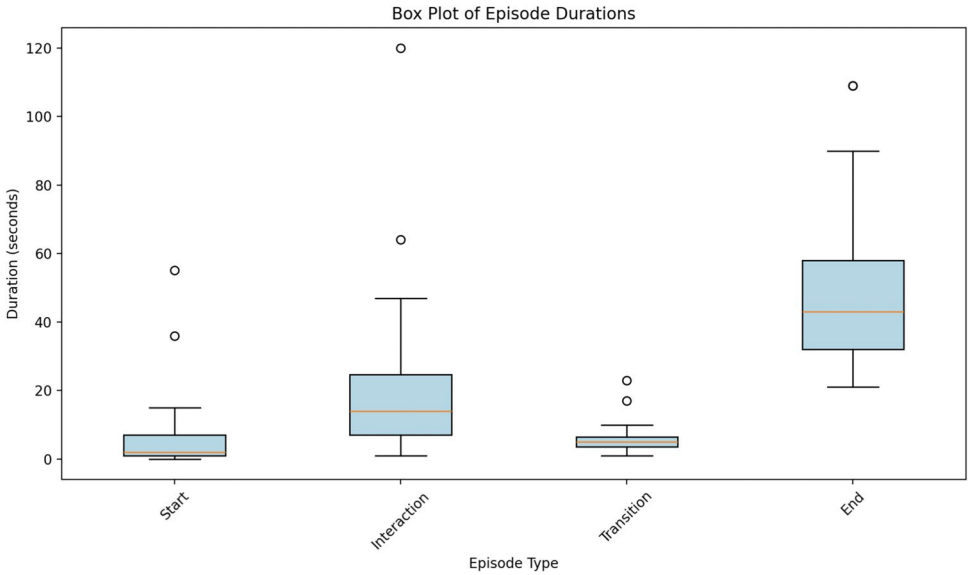

Fig. 5 Boxplot showing the distribution of episode durations for each episode type. Two outliers are observed: one corresponds to a TD child who is highly talkative, leading to longer episodes, while the other represents a child with ASD who requires more clinician guidance, resulting in extended durations

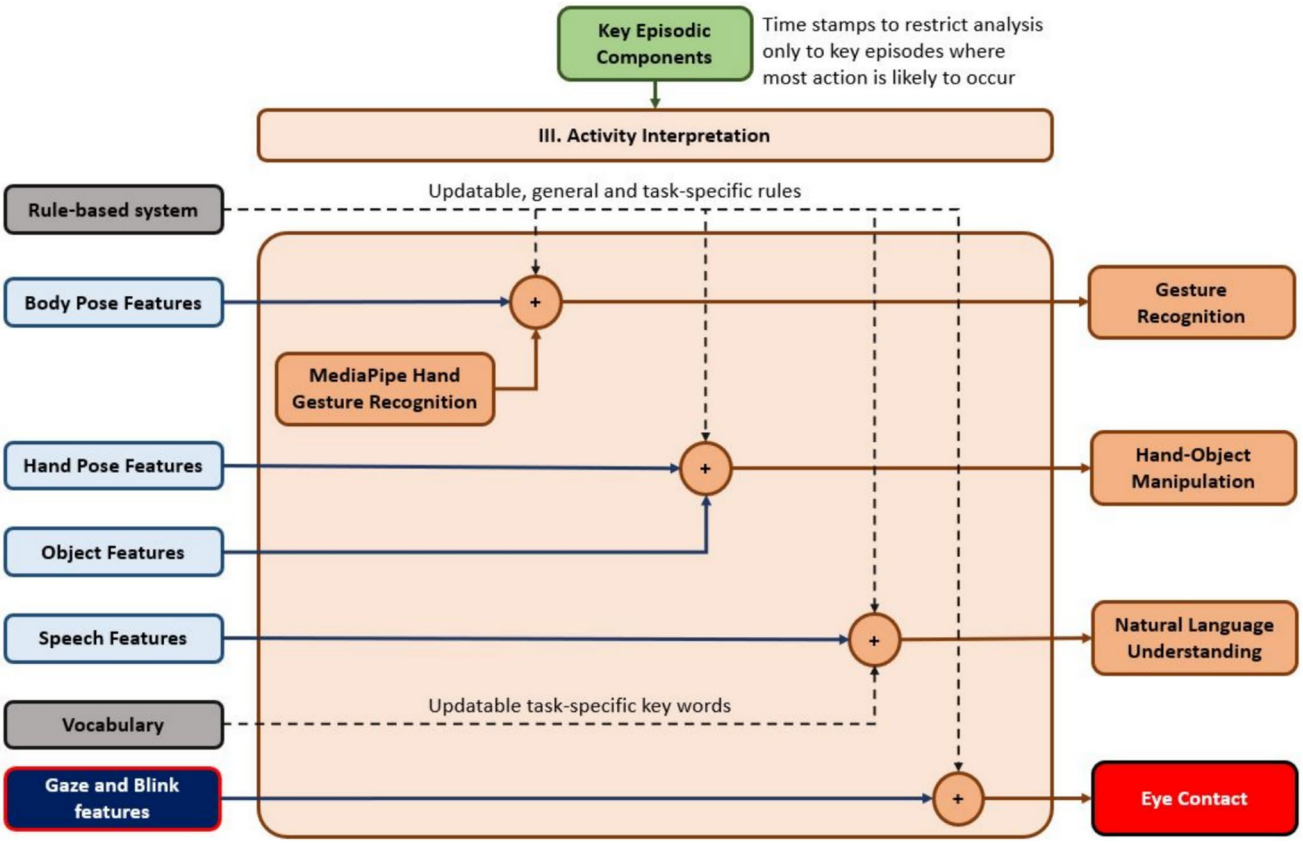

Fig. 6 Activity interpretation: within the detected key episodes, we analyze the scene by combining features using a set of updatable rules and classifiers, in order to identify gestures, hand-object manipulations and understand the natural language of the conversations. We incorporate new gaze and blink features for estimating eye contact

words ["I," "can," "have," "mind," "if," "piece," "pieces," "need," "more," "some," "please," "one," "two," "three," "four," "want," "ask," "few," "puzzle"] to recognize the intent behind the communication, which, in our case, is the request for more puzzle pieces.

Gesture Recognition

Detecting gestures is crucial, as they play a significant role in conveying non-verbal cues during human interactions and enable a deeper understanding of communication dynamics. We employ a combination of rules and classifiers, such as MediaPipe Hand. >Gesture Recognition [35], to detect and analyze these gestures. MediaPipe classifies the following hand gestures: Closed fist, Open palm, Pointing up, Thumbs down, Thumbs up, Victory, and Love. Within the detected request episode, typically between 2 and 6 s, we employ a 2-s window. If MediaPipe gesture recognition classifies at least a number of frames equal to one second (i.e., 30 frames) in the window as having a gesture, we consider the whole 2-s window as a single gesture. If gestures are the same across different windows, we treat them as one gesture. If the gestures differ, we use context from speech to decide the most appropriate gesture. For instance, a participant may verbally request for two pieces while showing the number two with their fingers. In those cases, we would select the "victory" label from MediaPipe's gestures. In the absence of context, we choose the gesture with the most occurrences or the highest confidence score. By recognizing these gestures, our system can offer some insights into human communication, helping to improve interactions and understanding of social dynamics.

Hand-Object Manipulation

Hand-object interactions offer important context and insights into an individual's intentions and actions during social interactions. We utilize the previously extracted object detection, segmentation, and hand pose outputs to identify and localize the objects and determine the hands' position and orientation relative to these objects. For hand-object manipulation, we focus on reaching and grabbing motions. We set a proximity threshold, indicating the distance between the hand's center and the center of the object within a specified radius. By evaluating hand and finger movement, if the hand moves toward the object and the relative distance between the thumb and other fingers is changing, we mark it as a potential reaching out for an object movement. Afterwards, if the object's position changes synchronously with hand and finger movement, a grabbing motion is detected.

Eye Contact

Eye contact occurs when two individuals simultaneously gaze into each other's eyes or, for simplicity, face. We investigate clinician-participant eye contact by integrating Tobii Glasses' clinician gaze and the previously extracted participant gaze using L2CS method. By filtering frames where both participant and clinician gaze directions align, we pinpoint moments of mutual eye contact. This approach provides information of interpersonal dynamics and non-verbal communication cues through gaze behavior. Given such details, it is possible to identify the following gaze pattern types [49]:

- Share: the state where two agents are looking at the same object or space
- Mutual: the state where two agents are having eye contact, which is also commonly known as the state where people are "looking at each other"
- Single: the state where the agent is looking at his/her partner while the other is looking elsewhere
- Miss: the state where the agent is looking away while being gazed at by his/her partner
- Void: the state where gazes of two agents do not form any meaningful interaction

In this work, we focus on the Mutual gaze type, which may potentially characterize both ASD and TD participants.

Results

Verbal and Non-verbal Request Detection During the Construction Task of ADOS-2

After extracting the relevant features and identifying key episodes, particularly the transition/request episode, we constrain the analysis to the duration of the episode, avoiding false detections in unnecessary intervals. Multiple request episodes may be detected per session, each typically preceded or followed by an interaction episode. Generally, a detected request episode has a duration of 2–6 s. The detection is considered valid if the annotation of the actual request interval, typically spanning 2 s, is within the 2–6 s interval of the detected request. As the detection process of the activity recognition module is based on predefined rules rather than trained models, all videos were used for evaluation. Due to the small dataset size, we applied the same rule set across all videos.

Table 4 Confusion matrix of verbal requests. TP, true positive; TN, true negative; FP, false positive; FN, false negative. The false negatives are due to transcription errors from the Whisper ASR model

		Actual request	
		Verbal request	No request
Predicted request	Verbal request	TP: 38	FP: 0
	No request	FN: 5	TN: 2

Table 5 Confusion matrix of gesture requests. TP, true positive; TN, true negative; FP, false positive; FN, false negative. The false negatives are due to MediaPipe's inability to recognize gestures

		Actual request	
		Gesture request	No request
Predicted request	Gesture request	TP: 4	FP: 0
	No request	FN: 3	TN: 23

Our CAI framework demonstrates the capability to effectively detect both verbal and non-verbal requests. The following evaluation metrics accuracy (ACC), precision (PREC), recall (REC), and F-1 score given in Eqs. (3)–(6) were used:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$PREC = \frac{TP}{TP + FP} \tag{4}$$

$$REC = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{PREC \times REC}{PREC + REC} \tag{6}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Confusion matrices (see Tables 4 and 5) comparing our method's verbal and gesture request detection against the annotated ground truth are presented. It is noteworthy that our proposed CAI framework achieved zero false positives due to its episode segmentation module, which excludes regions of the input without valuable information for detecting requests. This not only reduces false detections but also minimizes computational resource requirements. However, limitations in the feature extraction process led to a few false negatives, which are analyzed in detail in the "Discussion" section.

Figures 7 and 8 illustrate a comparison between the ground truth and detected requests for each participant. The following results were achieved for verbal request detection: accuracy, 89%; precision, 100%; recall, 88%; $F-1$ score, 94%; and for gesture request detection, accuracy, 89%; precision, 100%; recall, 57%; $F-1$ score, 73%.

These results indicate that, with current technologies, detecting non-verbal requests, precisely gestures, is more challenging than detecting verbal requests.

While having a high accuracy (89%) for both verbal and gesture requests, there is a notable difference in the precision (100%) and recall (57%) of the gesture detection, which shows the difficulties of the model in recognizing complex gestures, particularly those performed by ASD individuals.

In the "Discussion" section, we elaborate on some of the reasons why the system fails to detect certain requests, as well as provide potential future solutions.

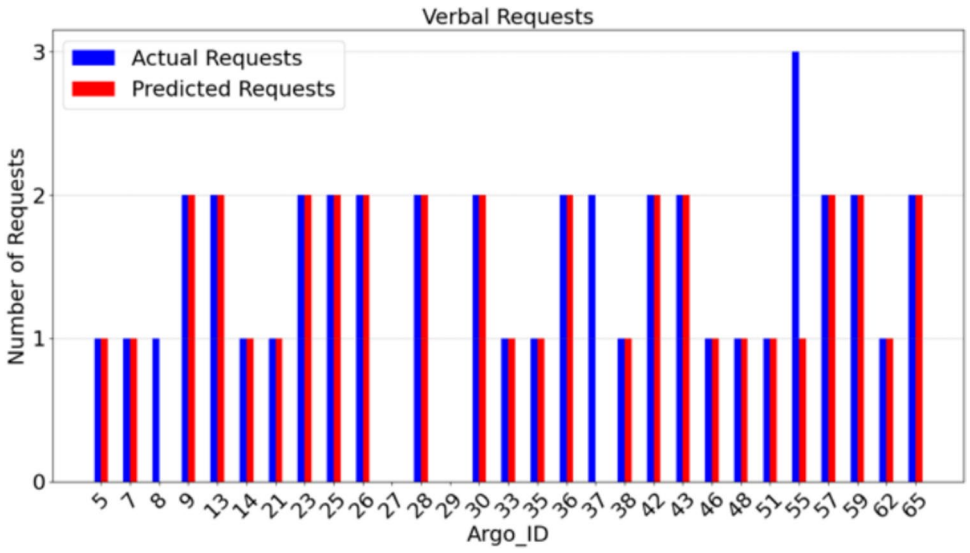

Fig. 7 Comparison of verbal requests between ground truth (GT) and detection (Pred) for 29 different participant IDs

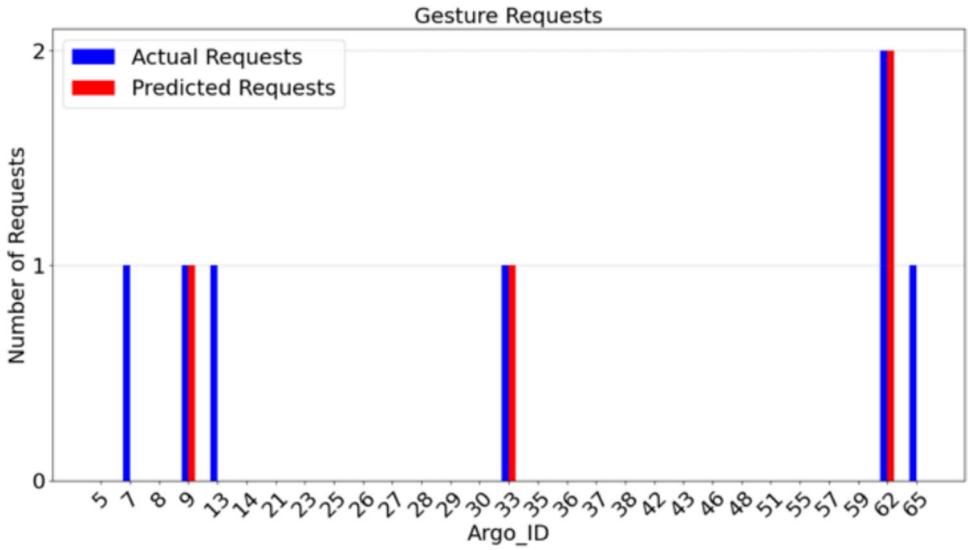

Fig. 8 Comparison of gesture requests between ground truth (GT) and detection (Pred) for 29 different participant IDs

Gaze and Blinking Statistics

The analysis of gaze and blinking behaviors provides valuable insights into socio-communicative patterns of people. In this subsection, we present the detected gaze, blinking and eye contact events during intervals when participants, either verbally or through gestures, requested additional puzzle pieces. In our context, gaze events refer to instances when participants looked at the clinician, regardless of whether the clinician reciprocated the gaze. On the other hand, eye contact events indicate mutual gaze, where both the participant and the clinician looked at each other.

The analysis focuses on request intervals for the following reasons:

- Faces of the participants are not consistently detected throughout the full Construction Task (due to challenges discussed in the "Discussion" section), making it unreliable for a comprehensive analysis. However, faces are detectable in a majority of the request intervals; thus, insights during such intervals can be derived.
- Making requests involves social interactions. By examining these intervals, we aim to better understand and compare the social cues exhibited by ASD and TD participants. This allows us to potentially uncover differences in how these cues are expressed between the two groups.

To ensure a reliable analysis of gaze and blinking, a preprocessing step was done, involving the exclusion of participant data where the face was not detected or detected for only a short period. Specifically, participants with a detected face ratio lower than 0.5 were excluded from the analysis due to the unreliability of drawing conclusions from limited data. Consequently, 5 ASD participants were removed, leaving the analysis with 16 ASD and 6 TD participants. For a summarized overview of the gaze and blinking events during the request intervals, see Table 6.

For gaze data, we analyzed and measured the total number and duration of participants' gazes toward the clinician during the Construction Task while making requests for more puzzle pieces.

Figure 9a shows that 9 out of 16 ASD and 4 out of 6 TD participants gazed toward the clinician. An outlier can be observed among the ASD participants, who gazed five times toward the clinician, in comparison to the average of one gaze exhibited by most of the other ASD participants.

Figure 9b presents the average gaze duration normalized by the number of gazes. Notably, a TD participant

Table 6 Summary of gaze and blinking events during request intervals when the face was visible. Total duration of detected faces (s) [request ratio] indicates the total time a face was detected and [the percentage relative to request duration]. Gaze duration [detected face ratio] indicates the total time spent looking at the clinician and [the percentage relative to detected face duration]. Blink metrics are described similarly

Group	# participants	# requests	Total duration of requests (s)	Total duration of detected faces (s) [request ratio]	Total gazes toward clinician	Total gaze duration (s) [detected face ratio]	Total blinks	Total blink duration (s) [detected face ratio]
ASD	21	35	130.0	82.1 [63%]	16	6.2 [8%]	13	1.2 [2%]
TD	6	12	45.0	35.1 [78%]	11	4.8 [14%]	11	1.4 [4%]

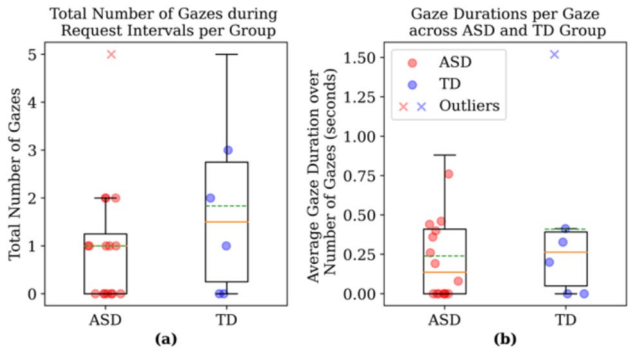

Fig. 9 **a** Box plot for total number of gazes per group, alongside the data points. Nine out of 16 ASD and 4 out of 6 TD participants were gazing toward the clinician. An outlier can be observed among the ASD participants. **b** Box plot for average gaze duration over number of gazes per group, alongside the data points. No outliers are observed.

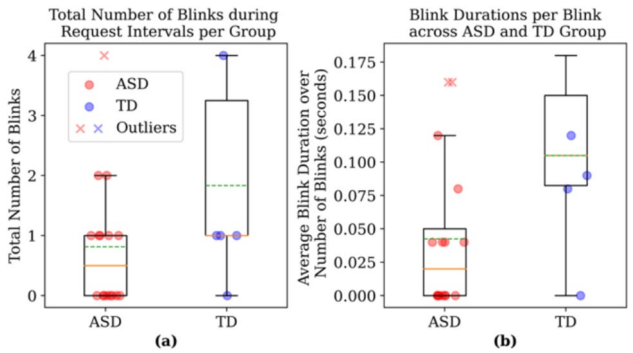

Fig. 10 **a** Box plot for total number of blinks per group, alongside the data points. Eight out of 16 ASD and 5 out of 6 TD participants were blinking. An outlier can be observed among the ASD participants. **b** Box plot for average blink duration over number of blinks per group, alongside the data points. Outliers can be observed among the ASD participants.

stands out as an outlier with a gaze duration of 1.50 s, significantly larger than the median of 0.24 s.

Regarding blink data, we analyzed and measured the total number and duration of blinks during the Construction Task while making requests for more puzzle pieces. Figure 10a displays that 8 out of 16 ASD and 5 out of 6 TD participants exhibited blinking. An outlier can be observed among the ASD participants, who blinked four times, in comparison to the average of approximately one blink by the other ASD participants. Figure 10b presents the average blink duration normalized by the number of blinks. The figure offers insights about the average duration of each blink. Some outliers are observed within the ASD participants. Due to reasons discussed in the "Discussion" section, in some instances, the blink was detected in a single frame, thus the duration is relatively short.

Considering the gaze information obtained from both participants and the clinician, we also analyzed eye contact events. The eye contact analysis is refined with blink information. During blinking intervals, gaze vector estimations may become inaccurate. Therefore, utilizing blinking intervals allows us to filter out imprecise gaze estimations.

Due to the short interval periods of the requests, any instance in which a participant and clinician looked at each other is considered an eye contact, regardless of its duration. This approach ensures that even subtle instances of eye contact are included in the analysis.

From our results, we observed that in almost all cases where a participant directed their gaze toward the clinician, it resulted in an eye contact event. This observation can be attributed to the continuous monitoring and assessment performed by the clinician during the task, leading to consistent efforts to establish and maintain eye contact. Consequently, the resulting boxplots for eye contact closely resemble those for gaze.

The results, despite limitations such as intermittent face detection and small intervals, indicate that gaze and blinking events can be detected, providing valuable insights.

However, given the small sample size, we refrain from drawing conclusive statements. Instead, we emphasize characterizing each participant to offer additional information that may be overlooked by clinicians.

The diverse gaze and blinking behaviors observed among participants underscore the need to characterize each individual in the assessment test, demonstrating a deeper understanding of within-group variability and identification of unique patterns, outliers, and potentially essential markers for recognizing abnormal gaze and blinking patterns.

This characterization can serve as a basis for personalized assessments or treatments catering to the specific needs of each participant.

Discussion

Verbal Requests

Our CAI framework uses speech features together with rules to detect verbal requests. Transcripts and speaker diarization information are used in combination with a dictionary of common request words, which successfully detects 38 out of 43 verbal requests, avoiding any false positives, mainly due to the episodic approach used in our system.

However, the performance of verbal request detection strongly depends on the accuracy of transcripts generated

by Whisper. Using the most accurate Whisper model, it fails to detect five requests due to the following reasons:

- Inaudible requests: One of the ASD participants barely speaks or does so very quietly. Additionally, with a suboptimal audio quality, it becomes challenging for even a state-of-the-art model such as Whisper to transcribe what was said.
- Transcript errors: For a particular participant, Whisper failed to transcribe the keywords "more," "two," which are crucial for the request, by instead transcribing "or" and "to." Without such keywords, the vocabulary could not be used to find their occurrences, in order to detect the request.
- Other factors: For the remaining cases, Whisper did not transcribe the keywords, possibly due to fast-paced speech or speech impairments. Such impairments can occur, particularly with participants high in the autism spectrum [50].

In the future, improving audio clarity through noise reduction and speech amplification could enhance request detection, particularly for low-volume speech. Additionally, large language models (LLMs) could refine transcriptions by correcting diarization [51] and keyword errors, as well as potentially be used to detect verbal requests given the proper context and prompt. Phoneme-based models may help recognize fast-paced speech and adapting ASR models to ASD-specific speech patterns could further improve transcription accuracy in clinical settings.

Non-verbal Requests: Gestures

Non-verbal communication, particularly through gestures, is vital for human-to-human interaction.

While our CAI framework can detect various gestures, challenges arise in specific instances. Figure 11 illustrates instances where our CAI framework encountered difficulties in detecting gestures, particularly among individuals with ASD, who may express gestures with great variability and with potential deficits [52].

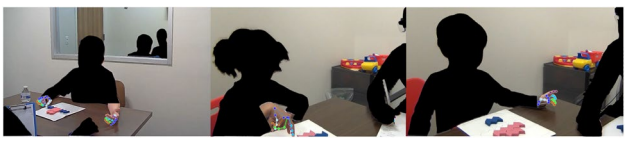

Fig. 11 Examples of complex, pragmatic and context-dependent gestures beyond the capabilities of our methods, thus not detected by our CAI framework. **a** The participant is expressing the need for more puzzle pieces and prepares to receive them, by combining "pointing" and "palm out" gestures. **b** The participant is expressing the need for two additional puzzle pieces on a specific position, by combining "pointing" and "number two" gestures. **c** The participant is expressing the need for more puzzle pieces and prepares to receive them, by combining "pointing" and "palm out" gestures

Most challenges were due to combinations of gestures resulting in complex and pragmatic gestures, making it difficult for the system to detect them.Some gestures are context-dependent, exceeding the capabilities of our limited set of rules combined with MediaPipe. Our system can successfully detect the "number two" gesture and pointing individually, but fails when an ASD participant combines both gestures to indicate needing two puzzle pieces at a specific location by pointing with two fingers. The unique expression of the number two, upside down with minimal distance between index and middle fingers, posed challenges for detection (refer to Fig. 11b).

Other examples involve participants combining pointing with a potential reaching out gesture, signaling a desire for more puzzle pieces with an open hand to receive them (refer to Figs. 10a and 11c).

In the future, expanding training data with ASD-specific and diverse gestures, including combination and context-dependent gestures, could enhance recognition accuracy. Implementing multi-label classification would improve detection of simultaneous gestures, while data augmentation can help the model recognize gestures across different orientations. Additionally, integrating temporal modeling could enable better interpretation of dynamic gesture sequences, making the system more responsive to nuanced communication cues.

Non-verbal Cues: Gaze, Blinking, and Eye Contact

Understanding non-verbal social cues offers insights into individuals' behaviors and internal states [20]. Manual tracking of these cues is challenging, and without invasive methods or specialized hardware, estimation becomes difficult. In our study, Tobii eye glasses provided gaze information for clinicians, and its camera feed was to estimate gaze and blinking from participants.

During the Construction Task, participants were often observed looking down while working on the puzzle, limiting the visibility of their faces. Consequently, the number of detected faces and subsequent blink and gaze estimations were affected. The average pitch angle for both ASD and TD groups indicated a downward head direction, further complicating the detection process. The detected face ratio given by the number of frames with a detected face confidence score larger than 0.5, divided by total number of frames, averaged around 61% for both groups throughout the entire Construction Task.

Challenges also arose due to the Tobii camera's fixed position on the clinician's face. In some instances, clinicians were not looking at the participant because of taking notes or picking up an object, and in other cases, the

children might have been outside the camera's field of view. Furthermore, camera motion resulted in significant blur, complicating the detection of faces and the estimation of gaze and blink behavior.

Examining gaze behaviors, we observed considerable variability among participants. Notably, the presence of outliers in both the ASD and TD groups suggests that individual differences play a significant role in socio-communicative responses. One participant with ASD exhibited an unusually high frequency of gazes toward the clinician, which further shows the variability within the spectrum of autism. Conversely, in the TD group, an outlier with an extended gaze duration might signify a particularly focused or expressive response. These outliers highlight the diverse nature of social interactions and the importance of considering individual tendencies when interpreting gaze behaviors.

In terms of blinking, the presence of outliers lead us to consider about the potential factors influencing blink patterns. For instance, an ASD participant with a higher-than-average blink frequency might be indicative of individual variations in sensory processing or attention modulation [20]. Understanding the underlying factors contributing to these outliers is crucial for clinicians in contextualizing socio-communicative behaviors.

Despite the limitations and challenges encountered, we were able to detect gaze, blink, and eye contact events during the short request intervals with high confidence. This is promising for clinicians, as it allows us to capture relevant social cues that may be missed during multitasking. Recognizing the potential impact on assessments, our focus is on providing supplementary information to enhance the diagnostic process rather than diagnosing ASD. However, we acknowledge that the small sample size, with only 27 usable Tobii videos, limits the ability to draw statistically significant conclusions, emphasizing the preliminary nature of our observations. This small dataset may lead to an over-representation or underrepresentation of certain behaviors or request types, which could limit the model's generalizability across different populations.

Furthermore, the difference in the number of ASD (20 participants) versus TD (7 participants) participants in the dataset could introduce biases in the detection of social cues, potentially affecting the model's ability to generalize across different subgroups. Given these factors, we focused on characterizing each participant individually, acknowledging the significant variability within both the ASD and TD groups, rather than comparing gaze and blinking behaviors between these groups.

To mitigate these biases, future work will involve expanding the dataset to include a more diverse and larger sample, which will help improve the model's performance and its ability to generalize across different demographics, languages, and social contexts. We believe that such advancements will enhance the robustness and applicability of our CAI framework, ultimately providing a more reliable tool for clinical use.

Conclusion

In summary, our framework presents a promising tool for clinicians, rather than a standalone diagnostic solution. It aims to assist in analyzing and understanding autism spectrum disorder (ASD) behaviors by offering detailed insights into social interactions. By incorporating deep learning algorithms and a rule-based system, our framework automates the analysis of the Construction Task of ADOS-2, providing clinicians with enhanced insights into verbal and non-verbal requests. Specifically, our CAI framework enhances time efficiency by automatically detecting key events, allowing clinicians to focus on critical moments instead of manually reviewing entire sessions. Furthermore, it supports clinician decision-making by detecting subtle communication cues like gaze and blinking, which are difficult for clinicians to accurately quantify, ultimately providing data that aids in making more informed diagnostic decisions.

Our CAI framework has demonstrated robust performance in detecting verbal requests, achieving a 94% $F-1$ score, with zero false positives and minimal false negatives due to limitations of the Whisper model. However, challenges arose in the detection of non-verbal requests, particularly complex, pragmatic, and context-dependent gestures performed by individuals with ASD, from which the rule-based system and MediaPipe could not detect, resulting in an $F-1$ score of 73%. The incorporation of Tobii glasses as a novel camera source has allowed us to estimate non-verbal cues such as gaze, blinking, and eye contact. Examination of these cues reveals variability among participants, with outliers suggesting that individual differences play a significant role in socio-communicative responses and underscore the importance of considering individual tendencies when interpreting behaviors, paving the way for more tailored and effective strategies in autism research and support.

In the future, with an expanded dataset, exploring the potential for classifiers or clustering methods may offer insights into distinguishing between ASD and TD, based on detected social cues. Our CAI framework as an ecological solution holds promise for adoption in clinical environments, providing a more comprehensive understanding of social interactions and behaviors during specific intervals and automating processes which are rather manual and time consuming.

Author Contribution BC.DSM: Conceptualization, Methodology, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. K.K: Software, Formal Analysis, Methodology, Investigation, Data Curation, Writing – Review & Editing. A.F: Software, Writing – Review & Editing. L.X: Software, Data Curation, Writing – Review & Editing. V.V: Software, Writing – Review & Editing. L.S: Data Curation, Writing – Review & Editing. E.D: Writing – Review & Editing. P.K: Software, Writing – Review & Editing. A.S: Software, Writing – Review & Editing. M.C Writing – Review & Editing. K.F: Conceptualization, Supervision, Writing – Review & Editing. A.L: Conceptualization, Supervision, Funding acquisition, Writing – Review & Editing.

Funding Open access funding provided by Eötvös Loránd University. This work was supported by the European Union project RRF-2.3.1–21-2022–00004 within the framework of the Artificial Intelligence National Laboratory and from the European Union's Horizon 2020 research and innovation program under grant agreement no. 952026.

Data Availability The data that support the findings of this study are not openly available due to reasons of sensitivity.

L.S receives consulting fees and research support from F. Hoffman- La Roche, royalties from Hogrefe Public and has equity interest in Argus Cognitive Inc. A.L has equity interest in Argus Cognitive Inc. M.C is a member of the Editorial Board of the Cognitive Computation Journal. A.S and P.K were employed by Argus Cognitive Inc while engaged in the research project. The remaining authors have no relevant financial or non-financial interests to disclose.

References

1. Maenner MJ, Warren Z, Williams AR, al et. 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. 2023;72:1–14.
2. Rehg JM. Behavior imaging and the study of autism. Proc 15th ACM Int Conf Multimodal Interact [Internet]. New York, NY, USA: Association for Computing Machinery; 2013. p. 1–2. Available from: https://doi.org/10.1145/2522848.2532203
3. Arac A, Zhao P, Dobkin BH, Carmichael ST, Golshani P. DeepBehavior: a deep learning toolbox for automated analysis of animal and human behavior imaging data. Front Syst Neurosci [Internet]. 2019;13. Available from: https://www.frontiersin.org/articles/https://doi.org/10.3389/fnsys.2019.00020
4. Ouss L, Palestra G, Saint-Georges C, Leitgel Gille M, Afshar M, Pellerin H, et al. Behavior and interaction imaging at 9 months of age predict autism/intellectual disability in high-risk infants with West syndrome. Transl Psychiatry. 2020;10:54.
5. Anagnostopoulou P, Alexandropoulou V, Lorentzou G, Lykothanasi A, Ntaountaki P, Drigas A. Artificial intelligence in autism assessment. Int J Emerg Technol Learn IJET. 2020;15:95–107.
6. de Belen RAJ, Bednarz T, Sowmya A, Del Favero D. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. Transl Psychiatry. 2020;10:333.
7. Drimalla H, Landwehr N, Baskow I, Behnia B, Roepke S, Dziobek I, et al. Detecting autism by analyzing a simulated social interaction. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G, editors., et al., Mach Learn Knowl Discov Databases. Cham: Springer International Publishing; 2019. p. 193–208.
8. Gartner. 5 trends drive the Gartner hype cycle for emerging technologies, 2020 [Internet]. 2021 [cited 2024 Sep 1]. Available from: https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020.
9. Dos Santos Melicio BC, Xiang L, Dillon E, Soorya L, Chetouani M, Sarkany A, et al. Composite AI for behavior analysis in social interactions. Companion Publ 25th Int Conf Multimodal Interact [Internet]. New York, NY, USA: Association for Computing Machinery; 2023. p. 389–97. Available from: https://doi.org/10.1145/3610661.3616237
10. Freeth M, Morgan EJ. I see you, you see me: the impact of social presence on social interaction processes in autistic and non-autistic people. Philos Trans R Soc B Biol Sci. 2023;378.
11. Ghafghazi S, Carnett A, Neely L, Das A, Rad P. AI-augmented behavior analysis for children with developmental disabilities: building toward precision treatment. IEEE Syst Man Cybern Mag. 2021;7:4–12.
12. Lu J, Nguyen M, Yan WQ. Deep learning methods for human behavior recognition. In: 35th international conference on image and vision computing New Zealand (IVCNZ), vol. 2020. Wellington: IEEE; 2020. pp. 1–6.
13. Hu K, Jin J, Zheng F, Weng L, Ding Y. Overview of behavior recognition based on deep learning. Artif Intell Rev. 2023;56:1833–65.
14. Georgescu AL, Koehler JC, Weiske J, Vogeley K, Koutsouleris N, Falter-Wagner C. Machine learning to study social interaction difficulties in ASD. Front Robot AI [Internet]. 2019;6. Available from: https://www.frontiersin.org/articles/https://doi.org/10.3389/frobt.2019.00132
15. Ioannis Vogindroukas E-NC Margarita Stankova, Proedrou A. Language and speech characteristics in autism. Neuropsychiatr Dis Treat. 2022;18:2367–77.
16. Mohanta A, Mukherjee P, Mirtal VK. Acoustic features characterization of autism speech for automated detection and classification. Natl Conf Commun NCC. 2020;1–6.
17. Zhou H, Zhang Y, Hu H, Yan Y, Wang L, Lui SSY, et al. Neural correlates of audiovisual speech synchrony perception and its relationship with autistic traits. PsyCh J. 2023;12:514–23.
18. Celiktutan O, Wu W, Vogeley K, Georgescu AL. A computational approach for analysing autistic behaviour during dyadic interactions. Pattern Recognit Comput Vis Image Process ICPR 2022 Int Workshop Chall Montr QC Can August 21–25 2022 Proc Part I [Internet]. Berlin, Heidelberg: Springer-Verlag; 2023. p. 167–77. Available from: https://doi.org/10.1007/978-3-031-37660-3_12
19. Chi NA, Washington P, Kline A, Husic A, Hou C, He C, et al. Classifying autism from crowdsourced semistructured speech recordings: machine learning model comparison study. JMIR Pediatr Parent. 2022;5: e35406.
20. Goldberg TE, Maltz A, Bow JN, Karson CN, Leleszi JP. Blink rate abnormalities in autistic and mentally retarded children: relationship to dopaminergic activity. J Am Acad Child Adolesc Psychiatry. 1987;26:336–8.

21. Nakano T, Kato N, Kitazawa S. Lack of eyeblink entrainments in autism spectrum disorders. Neuropsychologia. 2011;49:2784–90.
22. Babu PRK, Aikat V, Martino JMD, others. Blink rate and facial orientation reveal distinctive patterns of attentional engagement in autistic toddlers: a digital phenotyping approach. Sci Rep. 2023;13:7158.
23. Madipakkam AR, Rothkirch M, Dziobek I, Sterzer P. Unconscious avoidance of eye contact in autism spectrum disorder. Sci Rep 2017;7:13378. Available from: https://doi.org/10.1038/s41598-017-13945-5.
24. Frazier TW, Strauss M, Klingemier EW, Zetzer EE, Hardan AY, Eng C, et al. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. J Am Acad Child Adolesc Psychiatry. 2017;56:546–55.
25. Guo Z, Kim K, Bhat A, Barmaki R. An automated mutual gaze detection framework for social behavior assessment in therapy for children with autism. Proc 2021 Int Conf Multimodal Interact [Internet]. New York, NY, USA: Association for Computing Machinery; 2021;444–52. Available from: https://doi.org/10.1145/3462244.3479882
26. Alvari G, Coviello L, Furlanello C. EYE-C: eye-contact robust detection and analysis during unconstrained child-therapist interactions in the clinical setting of autism spectrum disorders. Brain Sci. 2021;11:1555.
27. Wang CY, Bochkovskiy A, Liao HY. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023. pp. 7464–75.
28. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 [Internet]. 2023 [cited 2024 Sep 1]. Available from: https://github.com/ultralytics/ultralytics.
29. Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, Luo P, Liu W, Wang X. Bytetrack: multi-object tracking by associating every detection box. In: European conference on computer vision. Cham: Springer Nature Switzerland; 2022. pp. 1–21.
30. Sarandi I, Linder T, Arras KO, Leibe B. MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. IEEE Trans Biom Behav Identity Sci. 2021;3:16–30.
31. Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F, Grundmann M. Blazepose: on-device real-time body pose tracking. arXiv preprint arXiv:2006.10204. 2020;
32. Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang CL, Yong MG, Lee J, Chang WT. Mediapipe: a framework for building perception pipelines. arXiv preprint arXiv:1906.08172. 2019.
33. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision; 2017. pp. 2961–9.
34. Cheng HK, Tai YW, Tang CK. Modular interactive video object segmentation: interaction-to-mask, propagation and difference aware fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. pp. 5559–68.
35. Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang CL, Grundmann M. Mediapipe hands: on-device real-time hand tracking. arXiv preprint arXiv:2006.10214. 2020.
36. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. pp. 1290–9.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. pp. 770–8.
38. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. PMLR; 2023. pp. 28492–518.
39. Koluguri NR, Park T, Ginsburg B. Titanet: neural model for speaker representation with 1d depth-wise separable convolutions and global context. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2022. pp. 8102–6.
40. Serengil SI, Ozpinar A. HyperExtended LightFace: a facial attribute analysis framework. 2021 Int Conf Eng Emerg Technol ICEET [Internet]. IEEE; 2021. p. 1–4. Available from: https://doi.org/10.1109/ICEET53442.2021.9659697
41. Kuprashevich M, Mivolo TI. Multi-input transformer for age and gender estimation. In: International conference on analysis of images, social networks and texts. Cham: Springer Nature Switzerland; 2023. pp. 212–26.
42. Birkl R, Wofk D, Müller M. Midas v3. 1--a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460. 2023.
43. Fodor Á, Fenech K, Lőrincz A. BlinkLinMulT: transformer-based eye blink detection. J Imaging. 2023;9(10):196.
44. Grishchenko I, Ablavatski A, Kartynnik Y, Raveendran K, Grundmann M. Attention mesh: High-fidelity face mesh prediction in real-time. arXiv preprint arXiv:2006.10962. 2020;
45. Ablavatski A, Vakunov A, Grishchenko I, Raveendran K, Zhdanovich M. Real-time pupil tracking from monocular video for digital puppetry. arXiv preprint arXiv:2006.11341. 2020;
46. Guo J, Zhu X, Yang Y, Yang F, Lei Z, Li SZ. Towards fast, accurate and stable 3d dense face alignment. Eur Conf Comput Vis. Springer; 2020. p. 152–68.
47. Abdelrahman AA, Hempel T, Khalifa A, Al-Hamadi A, Dinges L. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In: 2023 8th international conference on frontiers of signal processing (ICFSP). IEEE; 2023. pp. 98–102.
48. McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M. Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Proc Interspeech 2017; 2017. pp. 498–502. https://doi.org/10.21437/Interspeech.2017-1386.
49. Chang F, Zeng J, Liu Q, Shan S. Gaze pattern recognition in dyadic communication. In: Proceedings of the 2023 symposium on eye tracking research and applications; 2023. pp. 1–7.
50. Mody M, Belliveau JW. Speech and language impairments in autism: insights from behavior and neuroimaging. North Am J Med Sci. 2013;5:157–61.
51. Wang Q, Huang Y, Zhao G, Clark E, Xia W, Liao H. Diarizationlm: speaker diarization post-processing with large language models. arXiv preprint arXiv:2401.03506. 2024.
52. Fourie E, Palser ER, Pokorny JJ, Neff M, Rivera SM. Neural processing and production of gesture in children and adolescents with autism spectrum disorder. Front Psychol [Internet]. 2020;10. Available from: https://www.frontiersin.org/articles/https://doi.org/10.3389/fpsyg.2019.03045