ASD-HI: A Parent-Child Interaction Dataset for Automated Assessment of Home Intervention

Zhaohui Li[1], Yusuf Akemoglu[1,2], Jincheng Lyu[3], Qingxiao Zheng[1], and Jinjun Xiong[1]

[1] Department of Computer Science and Engineering, University at Buffalo
{zli253,qingxiao,jinjun}@buffalo.edu
[2] Department of Special Education, Duzce University, Duzce, Turkey
yusufakemoglu@gmail.com
[3] The Pennsylvania State University jjl6992@psu.edu

Abstract. Home-based interventions are vital for supporting young children with autism spectrum disorder (ASD), yet many parents struggle to implement strategies effectively due to limited training. While specialists such as educators and speech-language pathologists provide guidance, real-time feedback outside professional settings remains scarce. To bridge this gap, we leverage advances in AI to support parents through automated assessment. However, training such AI systems requires robust data, which is currently limited. To address this, we created ASD-HI (Autism Spectrum Disorder - Home Intervention), a multi-modal dataset comparing 473 real parent-child interaction videos across three families. ASD-HI supports two core tasks: 1) Strategy Detection, identifying the behavioral strategies parents use, and 2) Fidelity Assessment, assessing the fidelity with which these strategies are implemented. We also propose a prompting-based LLM pipeline as a reference approach. It achieves 74% recall and 50% precision for strategy detection and 60% accuracy for fidelity assessment. Our work lays a foundation for developing AI-driven tools to enhance home interventions and improve outcomes for children with special needs.

Keywords: Autism Spectrum Disorder · Large Language Models · Home-based Intervention · Early Intervention · Special Education.

1 Introduction

Effective home-based interventions are critical in promoting the development of children with autism spectrum disorder (ASD), particularly within special education programs [19, 26]. Parent participation and parent-implemented home-based interventions have been used to promote the social-emotional, social communication, and cognitive development of children with ASD [24]. However, many parents face substantial challenges implementing recommended interventions at home due to time constraints and insufficient training [4]. At the same time, special education professionals often struggle to provide continuous support beyond the classroom due to resource limitations and large caseloads [9]. These barriers highlight a critical gap in sustaining effective home-based interventions.



Fig. 1: Examples of shared reading from the dataset, where parents use strategies (gestures, expressions, discussion) to engage children and build joint attention.

Educators widely use artificial intelligence (AI) tools like ChatGPT [20] to generate instructional materials, design interventions, and even conduct assessments in recent five years [14, 29]. In special education, AI has been applied to automate child engagement assessment [28], create individualized interventions [17], and monitor progress for children with ASD [31, 32]. Recent work also shows that LLMs can analyze ASD clinical conversations as well as, or better than, non-expert humans [12]. However, adoption in special education remains slower than in general education [16], due to the highly individualized nature of home-based interventions and the lack of high-quality data, compounded by ethical concerns in data collection [13].

To address these challenges, we introduce ASD-HI, a multi-modal dataset derived from real-world parent-child interaction videos (Fig. 1), designed to support AI-driven monitoring of home-based interventions. The data were originally collected during the i-PiECS (Internet-based Parent-implemented Early Communication Strategies-Storybook) project [1, 2]. i-PiECS trained parents in naturalistic communication strategies (NCT) during storybook reading with their children with ASD. Approved by IRB, in i-PiECS, ASD experts manually evaluated parent recordings and provided detailed feedback. Building on these annotations, ASD-HI automates the process via two tasks: *Strategy Detection*, identifying NCT strategies used, and *Fidelity Assessment*, evaluating how well they are implemented. To convert i-PiECS data into AI tasks, we aligned human-annotated transcripts with video using an automatic speech recognition (ASR) model to establish ground-truth time frames for Strategy Detection. Discrepancies were resolved through additional human annotation to ensure reliability. The final ASD-HI dataset includes 478 labeled instances (4.25 hours), split into 239 training, 120 validation, and 119 test samples.

We propose potential solutions for strategy detection and fidelity assessment, building on recent breakthroughs in AI and utilizing multi-modal learning techniques to integrate video, audio, and contextual data. To detect the NCT strategy, we first employed the Whisper ASR model [22] to obtain a transcript of the entire video. Next, we utilized JanusPro, a visual LLM [7], to generate a video description for each sentence in the transcript, allowing us to extract relevant

visual information. We then developed a greedy search algorithm (refer to Algorithm 1) to identify all parent strategies in the video, using expert-designed LLM prompts alongside the transcripts and video descriptions. We formulated fidelity assessment as a video classification problem and designed a classification prompt using LLMs, incorporating expert-designed prompts with the transcripts and video descriptions. Our experiments assessed various pre-trained baseline approaches, including SOTA video understanding and classification models. The results demonstrate that our method, which combines transcript analysis with video description classification using GPT-4o, achieves the best performance, yielding around a 60% F1 score and accuracy. In the discussion section, we conduct an error analysis to highlight the main challenge of this task: identifying joint attention between parent and child.

ASD-HI provides training data for AI algorithms to automate the assessment of home-based interventions, enhance parental engagement, and improve educational outcomes for children with special needs. We have two main contributions: We created ASD-HI, the first large-scale, multimodal video dataset focused on assessing parent-child interactions in home-based interventions for children with ASD. The dataset includes 478 video recordings of reading sessions involving autistic children and their parents, providing high-quality, reliably annotated data. This resource is invaluable for research in computer vision and special education, facilitating advancements in AI-driven behavior analysis. We developed efficient AI pipeline solutions that utilize multimodal LLMs to automate the identification and evaluation of NCT strategies used by parents and children. By bridging the gap between AI techniques and domain expertise through LLM prompting, this pipeline supports scalable home-based ASD interventions.

2 Related Works

Recent multi-modal datasets for ASD primarily focus on AI-driven diagnosis using computer vision and machine learning techniques. [11] and [10] provide image-based ASD datasets for prediction using AI models. [8] introduced a dataset containing 22 hours of social interaction videos for autism risk assessment, though it is not publicly available. The dataset proposed in [33] includes 1,837 video sequences for gesture analysis in ASD detection. [6] compiled a dataset of 82 children's videos, using deep learning-based head-related feature extraction for ASD diagnosis. [23] proposed a dataset with 2,467 videos, incorporating frame convolutional and attention map features for ASD classification.

Despite these advancements, there is a critical lack of datasets focusing on parent-child interactions in home-based ASD interventions. The Parent-Teacher Interaction Study [21] examines parental involvement in the education of individuals with ASD, highlighting the impact of stress and well-being, but does not include direct interaction recordings. The Engagnition Dataset [15] captures children's engagement in a serious game, incorporating physiological and behavioral data with usability assessments from parents. The DREAM dataset comprises over 3,000 robot-assisted therapy sessions, totaling more than 300 hours, to assess educational and therapeutic interventions for children with ASD,

though it primarily focuses on structured interactions with robots rather than human-human engagement. Most of these existing datasets emphasize child-centric behaviors, such as reactions to stimuli or motor tasks, rather than dyadic parent-child or teacher-child interactions. Additionally, studies such as [21] rely on questionnaire-based assessments rather than multi-modal recordings of real-world interactions.

There are only two video datasets that focus on parent-child interactions in ASD domain: 1. The Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Module 3 (ADOSMod3) dataset includes recordings used to assess social behaviors in verbally fluent individuals, comprising 170 people diagnosed with autism and 120 with non-ASD conditions [18]. 2. The Remote Natural Language Sampling (Remote-NLS) dataset [5] consists of 89 Zoom video recordings, each capturing a 15-minute interaction between a parent and their child, aged 4 to 7 years, with ASD. However, both of these datasets focus on clinical settings rather than home-based interventions, and they primarily target child behavior rather than parent-child interactions. To the best of our knowledge, the ASD-HI dataset is the first to provide video data of parent-child interactions specifically designed for home-based ASD interventions.

3 Dataset Creation

Since there is currently no video dataset designed explicitly for the automated evaluation of home-based interventions, and given that both parents and specialists require additional support in this area, we propose the ASD-HI dataset. This dataset features expert-labeled annotations intended to train AI algorithms that can help assess and provide feedback on parental intervention strategies. To achieve this, we have developed two tasks: Strategy Detection and Fidelity Assessment, to bridge the gap between advanced AI techniques and the expertise of special educators, aiming to facilitate research in AI-driven solutions for special education and improve outcomes.

Fidelity	**Modeling**	**Mand-Model**	**Time Delay**	**Summary**
4	36	146	49	231
3	14	66	12	92
2	25	81	11	117
1	4	28	6	38
Summary	79	321	78	478

Table 1: Data statistics of fidelity scores across parent strategies

3.1 Backgound

The data for these two tasks were collected from real-world video recordings of parent-child interactions within the i-PiECS project, an evidence-based intervention program for children with ASD and their families [2, 1]. i-PiECS stands for internet-based parent-implemented early communication interventions storybook. In the i-PiECS program, parents are trained and coached to implement NCT strategies during storybook reading sessions with their children with ASD, fostering developmental progress through structured, clinically validated

evidence-based strategies. Across the i-PiECS studies, data were primarily collected through Zoom videoconferencing, where parents recorded their shared reading sessions using NCT strategies. ASD-HI focuses on three key NCT strategies: modeling, mand-model, and time delay. **Modeling** is a teaching strategy where the parent models words, phrases, or gestures, expecting the child to imitate. For example, the parent might label an object by saying *"Blue ball!"* while pointing to a picture of a blue ball, encouraging the child to respond by repeating. **Mand-Model** is similar but includes a verbal prompt in the form of a question, choice, or directive. For instance, the parent may ask, *"Is this an apple or a banana?"* or instruct *"Say 'more please.'"* Unlike modeling, the mand-model approach explicitly requires a response. Simply labeling an object expecting imitation (e.g., *"Ball."*) or asking a yes/no question (e.g., *"Is this a gorilla?"*) does not fit this strategy. **Time-Delay** encourages the child to initiate communication by pausing within familiar routines or activities. For example, while reading a book, a parent may leave a sentence incomplete, such as *"Edwin dropped one large box of ___,"* and wait expectantly for the child to complete it. Similarly, during a turn-taking activity, the parent might hold a page without turning it, waiting for the child to say *"My turn!"* Each of these strategies distinctly fosters language development by shaping how children engage in communication.

3.2 Task Design

The first task in the ASD-HI dataset is **Strategy Detection**. This task serves as a foundational step, without accurate segmentation, subsequent evaluation and feedback mechanisms cannot be reliably applied. We labeled 478 parent strategy uses with *start-time*, *end-time*, and *strategy label*. The strategy label has three classes of NCT strategies we mentioned before: *Modeling*, *Mand-Model*, and *Time Delay*. The goal of the strategy detection task is to identify the use of NCT strategies in full-length reading sessions. The output consists of the session *start-time*, *end-time*, and *strategy label*. We provide a dataset from three families, with the objective of detecting all 478 parent strategy uses in 48 reading sessions, including their respective start-time, end-time, and strategy type. To evaluate performance, we propose two metrics: *Coverage*: Measures the proportion of correctly detected NCT strategies used among all labeled data; *Accuracy*: Assesses how precise the detections are.

After identifying all strategy use instances, another AI algorithm is required to evaluate how effectively parents executed the NCT strategies. This evaluation is represented by a fidelity score. To achieve this, we introduce a second task, **Fidelity Assessment**, formulating it as a four-way classification problem. For ease of machine learning model training, we separate this task from the strategy detection task. We utilize the 478 labeled parent strategy video clips from the strategy detection task (by segmenting the reading session videos by start and end time label), along with the ground truth fidelity scores provided in the i-PiECS annotations. The evaluation metrics for this task follow standard machine learning classification practices, utilizing the F1 score and Accuracy.

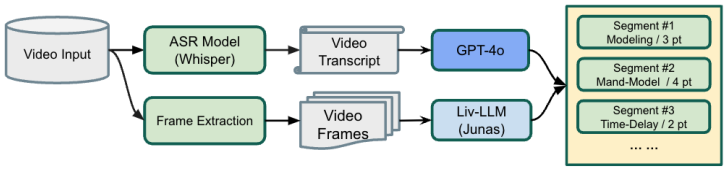


Fig. 2: Multi-modal pipeline for strategy detection.

3.3 Data Annotation

Parents in i-PiECS completed online training modules and received telepractice coaching via videoconferencing. It used a single-case multiple-baseline design and a 4-point fidelity scale to evaluate strategy implementation. Parent-child reading sessions were video recorded, and trained raters annotated strategy use and child responses by timestamp. Fidelity for modeling, mand-model, and time-delay strategies was rated from 1 (low) to 4 (high) based on criteria from [3], such as establishing joint attention, presenting the strategy, pausing, and giving feedback. Scores decreased if key steps were missed.

The original i-PiECS annotation has high reliability, primary raters annotated all assigned parent-child interaction videos by manually coding NCT strategies and child communication, while blinded secondary raters independently coded a subset, ensuring inter-rater reliability through discussions and achieving over 90% agreement. To evaluate the data reliability, the ASD-HI data annotation process begins with raw video footage paired with expert human-coded transcripts from i-PiECS. We leveraged ASR models to align i-PiECS transcripts with video timelines, segmenting recordings into shorter clips. Each clip corresponds to a NCT strategy, labeled with the specific communication strategy employed and a fidelity score reflecting adherence to i-PiECS protocols. During this step, we identified several missing lines in the i-PiECS annotated transcripts compared to ASR-generated transcripts, indicating potential inconsistencies in the original human-coded data. In addition, we conducted an additional round of human annotation. A trained computer scientist compared human annotations with machine-generated labels. If inconsistencies were detected, a separate ASD researcher reviewed and finalized the labeling. There are two main reasons for incorporating algorithmic checks in transcript validation and video segmentation: First, by using ASR models, we can assess the consistency and accuracy of human-coded transcripts from i-PiECS. If discrepancies exist, ASR can help identify them. Second, The i-PiECS dataset does not pre-segment reading session videos into labeled clips. Since video annotation is time-intensive, a hybrid approach where machine-generated labels are first applied and then verified by humans, optimizes efficiency while maintaining annotation accuracy. The data for each task is split into approximately 50% training, 25% validation, and 25% test sets. Detailed coding protocols and the dataset are available on our website[4].

[4] https://asdhi.xlabub.com/

4 Baseline Approach

In ASD-HI, there are two main AI challenges: 1) How can we identify strategy-specific segments within lengthy videos? This task is challenging because these segments often lack dramatic scene changes or clear verbal cues that indicate a transition. 2) How can we evaluate the fidelity of the strategies used by parents? This requires analyzing joint attention between the parent and child. However, there is no visual model that effectively captures this subtle visual pattern.

To tackle the strategy detection task, we first examined the field of video shot detection, which aims to identify and annotate distinct segments within a video by detecting abrupt shot changes [25]. The objective is to automatically pinpoint moments in a video where visual or thematic content shifts significantly, indicating the start or end of a shot. However, in our context, there are no substantial scene changes between strategies, rendering existing SOTA methods ineffective. In addition, the detection of strategy mainly relies on verbal conversation information. Consequently, we devised a baseline approach that leverages cues from the transcript and joint attention signals in the visual data. We propose a pipeline for integrated strategy detection that utilizes ASR and multi-modal LLM prompting. The pipeline comprises two stages: *Preprocessing* and *Processing* (details are provided in Algorithm 1).

In the preprocessing stage, we extract audio from the input videos to generate transcripts $T = \{s_1, s_2, \ldots, s_{|T|}\}$, where each sentence s_i is associated with a start time and an end time. This transcription is performed by the pre-trained Whisper model [22]. Meanwhile, video frames are sampled at a rate of 1 frame per second for visual analysis. Next, we employ Janus-Pro-1B [7] to generate visual descriptions for each frame by a prompt designed by ASD exert (see table 2), $V = \{v_1, v_2, \ldots, v_{|V|}\}$. We deliberately avoid using end-to-end video understanding models (e.g., VideoLLaMA [30]) for two reasons: (1) current models struggle to capture fine-grained interaction details essential for our task, such as parent-child joint attention; and (2) the recorded sessions often exceed two minutes in duration, making it impractical to generate a single comprehensive description to check joint attention. We chose Janus-Pro-1B for its advanced performance, compact model size, and open-source availability.

In the processing stage, we apply a greedy search algorithm (Algorithm 1) to identify all *strategy groups*, i.e., sequences of sentences representing an NCT strategy instance. Typically, a strategy group comprises three main sentences: a *parenting order*, a *child response*, and *parental feedback*, although additional sentences may also be included. The greedy search operates by scanning from the first sentence; once it detects a clue sentence indicating the start of a strategy group, it continues checking subsequent sentences to see if they belong to the same group. When it encounters a sentence that does not fit, the search terminates that group and resumes from the new sentence. This approach reduces computational complexity while maintaining robust accuracy.

Recognizing that fidelity assessment is a video classification task, we reviewed several SOTA classification models as baselines, including Transformer-based classifiers and pre-trained LLMs, to evaluate performance across different modal-

Prompt A: Generate an Image Description for Video Frame

<image_placeholder>
This image is from a parent-child interaction video. Please provide a detailed description of the image, focusing on the following aspects:
• **Parent's Actions:** Describe what the parent is doing, including any notable gestures or expressions.
• **Child's Reactions:** Detail the child's behavior or reactions.
• **Joint Attention:** Identify whether the parent and child are looking at the same object. Provide evidence from the image, then explicitly state either:
  - `*Yes, they have joint attention.*`
  - `*No, they do not have joint attention.*`
• `Visual Cues: Books or toys and other visual elements present in the frame.`

Prompt B: Summarize Video Descriptions from Frame Image Descriptions

You are provided with a series of detailed frame descriptions from a parent-child interaction video. Each frame description focuses on the following key elements: **Parent's Actions:**, **Child's Reactions: Joint Attention:**, **Visual Cues:** Books or toys and other visual elements present in the frame.
Your task is to combine these individual frame descriptions into a single comprehensive summary. Present whether the parent and child are looking at the same object. Provide evidence from the descriptions, then explicitly state either:
  - `'Yes, they have joint attention.'`
  - `'No, they do not have joint attention.'`

Please respond **only** in the following format (without additional commentary):
1. **Reasoning**: <your reasoning based on the evidence>
2. **Fidelity**: `'Yes, they have joint attention.'` | `'No, they do not have joint attention.'`
Frame Descriptions:
`<Concatenated frame descriptions here>`

Table 2: Prompts for describing parent-child interaction videos and images.

ities. However, as the results shown in the section (5), these methods did not perform well due to the lack of pre-trained parent-child interaction data. We decided to use the few-shot prompting classification method from the LLM, as described in Table 3 prompt D. We maintained the same transcript and video description that were used in the strategy detection task. This approach was chosen for its ease of use and the extensive pre-training it has undergone across various knowledge domains. Additionally, integrating fidelity assessment into the existing strategy detection pipeline allows for a seamless, fully automated system that performs both tasks simultaneously, as illustrated in Figure 2.

5 Experiments

This section presents the results of our baseline approach on the ASD-HI evaluation set. We conducted two sets of experiments to evaluate our baseline methods. First, we assessed our strategy detection pipeline through an ablation study using two different modality approaches. Then, we evaluated our prompt-based fidelity assessment method against SOTA baseline models.

For strategy detection, we designed our pipeline as described in Section 4. It incorporates two modality approaches: *Transcript-only* and *Video Description+Transcript*. The *Transcript-only* method relies solely on audio transcripts to identify the strategy sentences, while the *Video Description+Transcript* method combines both the transcript and video descriptions. The video-only approach is not applicable here because it depends entirely on video descriptions without transcripts; however, identifying the strategy requires verbal communication. We utilize the GPT-4o API for classifying the strategy sentences, and Janus-Pro-1B

Algorithm 1 The Greedy Strategy Search Algorithm

Input: Transcript sentence set $T = \{s_1, s_2, \ldots, s_{|T|}\}$, every s has an start time and end time
Input: Video Frame description set $V = \{v_1, v_2, \ldots, v_{|V|}\}$
Initialize result set: $O \leftarrow \emptyset$, set $i \leftarrow 1$
while $i \leq |T|$ **do**
    Compute video description d_i for sentence s_i using LLM method α (Prompt A&B, Table 2):

$$d_i = \alpha\Big(v_{s_i[\text{start_time}]},\ v_{s_i[\text{start_time+1}]},\ \ldots,\ v_{s_i[\text{end_time-1}]}, v_{s_i[\text{end_time}]}\Big)$$

    Classify s_i into NCT strategies {Modeling, Mand-model, Time-delay, None}:

$$NCT(s_i) = \beta(s_i, d_i), \text{where } \beta \text{ is a LLM Classifier (Prompt C in Table 3)}$$

    Skip the sentence if it does not use any strategy:

$$\text{if } NCT(s_i) = \text{None}, \quad i \leftarrow i + 1, \quad \text{continue}$$

    If s_i is using a strategy, initialize a strategy group: $g \leftarrow [s_i]$, and set $j \leftarrow i + 1$.
    If $\theta(g, s_j)$ is **True**, append s_j to g and increment $j = j + 1$.
    Repeat step 9 until $\theta(g, s_j)$ returns **False**, and append $g = [s_i, \ldots, s_{j-1}]$ to O.
end while
return O

is employed to generate the video descriptions. In all prompts, the temperature is set to zero to ensure consistency and reproducibility.

Table 4 presents the results on the ASD-HI validation set. Both pipeline methods show medium-level performance. The results indicate that the Transcript-only method achieves the best performance, covering 73.68% of NCT strategies used in a reading session. This suggests that it effectively detects most strategy usages. However, its accuracy is only 50.21%, highlighting the need for more precise detection in future tasks. Comparing *Transcript-only* with *Video Description + Transcript* serves as an ablation experiment to assess the contributions of different modalities. The results support the hypothesis that strategy use primarily relies on verbal conversation. Incorporating additional visual information, such as gestures and background details, does not enhance performance within our experimental setup.

Strategy Detection Method	Coverage	Accuracy
Transcript-only	73.68%	50.21%
Video Description+Transcript	68.42%	39.39%

Table 4: Strategy detection performance by coverage and accuracy

For fidelity assessment task, we compared with SOTA video classification with our prompting method via GPT families. We employed several video classification models as baselines, including training a transformer-based classifier and video LLM prompting. Below, we provide an overview of the baselines used in

Prompt C: Classify the NCT Strategies with Transcript and Video Description

System Prompt:
You are an expert in classifying parent-child intervention strategies. Your task is to read a line of transcript and video description and categorize the NCT strategy into one of three strategies: 1) **Modeling** - Parent demonstrates a thing in the book by saying one or two words. - Modeling can't be a sentence longer than five words. It focuses only on meaningful noun words that represent a thing in the book. - **Examples**: - Referring to a ball picture in a book, the parent says, "Blue ball!" expecting the child to imitate. - ... - **Non-Examples**: - Asking, "What do you have?" (This is a Mand-Model.) ... ... 2) **Mand-model** - ... ... 3) **Time Delay** - ... ...
User Prompt:
Classify the following text according to the categories above:
`- Video Description: "{description}" - Transcript: "{text}"`
Respond only with "Modeling", "Mand-model", "Time Delay" or "None".

Prompt D: Classify NCT Strategy use into Fidelity Scores

You are a Speech-Language Pathologist evaluating a parent's use of the **Modeling** strategy in a parent-child interaction.
Fidelity Scoring (1 to 4):
Award +1 point for each of the following:
1. Presenting a verbal or gestural model; 2. Establishing joint attention; 3. Waiting ∼3 seconds for the child's response; 4. Providing verbal feedback that is **NOT** a simple "Yes/No" question Now, read the following **transcript**, **video description**, and **waiting time** to evaluate based on the fidelity criteria above.
Provided Information:
`{video_description} {transcript}`

Response Format:
Please respond **only** in the following format (without additional commentary):
1. **Reasoning**: `<your explanation>` 2. **Fidelity**: `<1 | 2 | 3 | 4>`

Table 3: Prompts with examples for classifying and evaluating NCT strategies. We show simplified prompts due to the space limitation.

our experiments: **VideoMAE**: A SOTA performance video classification model with video masked autoencoder on several benchmarks [27]. We trained this on trainset, it can only trains this masked autoencoders on video frames data. **Transcript**: This baseline follows a two-stage process. First, we use the whisper model to extract transcripts from the videos. These transcripts are then used to train a BERT-based classifier, enabling classification based on spoken content. **Video + Audio**: This baseline also follows a two-stage process. First, we use Video-LLaMA [30] to generate video descriptions. These descriptions are then combined with ASR transcripts to train a BERT-based classifier, enabling classification based on both vision and audio content. **Video-LLaMA**: The method utilizes a prompting approach with Video-LLaMA, a LLM designed for video question answering with SOTA performance. This model leverages advanced language understanding to analyze video content based on specific prompts.

For our GPT prompting method, we also employ Janus-Pro-1B to generate video descriptions. We design three classification prompts for three different strategies; here, we only show the Modeling prompt due to the space limitation (see Table 3. Additionally, we evaluate classification performance across different GPT models, including GPT-4, GPT-4o, and the latest GPT-o1. Furthermore, we investigate performance differences across different modalities, comparing Transcript-Only versus Video + Transcript conditions.

As shown by the F1 and Accuracy scores in Table 3, our prompting method using GPT models generally outperforms other SOTA classifiers, particularly

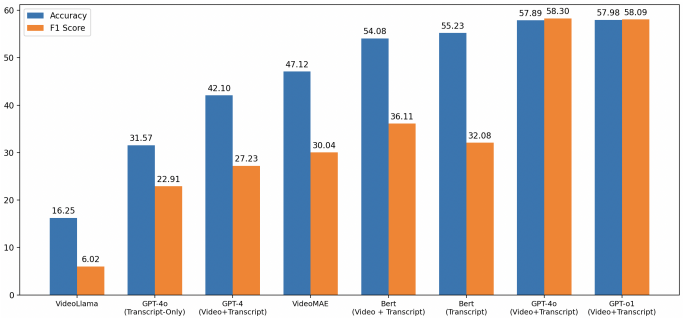


Fig. 3: Fidelity assessment: Accuracy and F1 by different baseline models

when compared to Video-LLaMA. The Video-LLaMA demonstrates poor performance, achieving only 16% accuracy. This could be attributed to the fact that the model is trained on different tasks and does not generalize well to our specific task. In contrast, text-based LLMs such as GPTs exhibit strong generalization capabilities. Among the various GPT prompting methods, using video and transcript together outperforms the Transcript-Only approach, as demonstrated in the GPT-4o experiments. Both GPT-4o and GPT-o1 (Video+Transcript) achieve the highest performance levels, exceeding the strongest baseline, Bert (Video+Transcript), by at least 4% in accuracy and 12% in F1 score. These results highlight the challenges of training a four-way video classification model for this task. A future direction is fine-tuning LLMs to improve performance.

In summary, our method demonstrates competitive performance in both Strategy Detection and Fidelity Assessment tasks. Notably, our prompting method achieves significant gains in fidelity classification compared to SOTA models. Since our approach does not require additional training, it can generalize well and is readily applicable to real-world scenarios. However, its effectiveness heavily depends on expert-designed prompts, and its performance still falls short of human-level accuracy. Future work should focus on fine-tuning LLMs or optimizing prompts to achieve further improvements.

6 Discussion and Conclusion

Even though our algorithm has achieved relatively high performance in both tasks, several challenges remain for future research. In strategy detection, our algorithm struggles with accurate identification, resulting in only 50% accuracy. This limitation arises from the inherent complexity of multimodal tasks and the error accumulation of greedy search. For example, if the algorithm incorrectly identifies a strategy that is not actually present, but that misidentified instance contains part of the next strategy usage sentence, it can lead to failures in subsequent detections. In fidelity assessment, our error analysis was driven by discrepancies between predicted labels and ground truth labels. We found that, in many cases, our algorithm systematically underestimates fidelity scores, often predicting values one point lower than the ground truth. For example, a data point with a ground truth fidelity score of 4 was frequently predicted as 3. The algorithm provided the following explanation: *"Reasoning: The parent used*

a question prompt and waited 3 seconds, then provided verbal feedback beyond yes/no. However, they did not consistently maintain joint attention. Fidelity: 3." Upon further inspection of the generated video description, we observed: *"Across most frames, the parent and child are shown looking at the same book, indicating joint attention while reading. However, in a later frame, the child focuses on a stuffed animal while the parent looks at the child, and in another frame, the child looks at the book while the parent looks at the child. These instances show that they are not always looking at the same object. No, they do not have joint attention."* The issue here is that the AI model rigidly evaluates joint attention across all video frames. If joint attention is absent in even a few frames, the AI fails to award the joint attention point. However, human raters are less strict and typically consider joint attention established if it occurs predominantly during the interaction. This error case highlights a fundamental challenge: controlling AI to apply human-like judgment criteria, especially when analyzing multi-modal data. Future work should focus on aligning AI evaluation strategies with human decision-making processes to reduce systematic discrepancies.

Our baseline approach has other limitations. First, we only tested the GPT family of LLMs, as its API is readily accessible and easy to use. Other LLMs require different formats and integrations, which we plan to explore in future work. Second, our prompt was designed by ASD experts, but we did not apply prompt optimization or fine-tune the model on the training set. Future research should investigate whether tailored fine-tuning or systematic prompt engineering could improve performance. Third, our model's accuracy remains limited, achieving only 50% in strategy detection, indicating the need for further refinement. Finally, dataset size is a constraint, limiting the ability to fully optimize our models. In fidelity classification, the LLM prompting method outperformed our trained classifier, suggesting that additional labeled data is necessary for improving model generalization. We will continue expanding and refining the dataset to establish a more robust resource for child behavior research.

Conclusion. We introduced ASD-HI, a new video dataset featuring real-world recordings of parent-child interactions during home-based interventions. The dataset includes accurately labeled data from human annotators, facilitating the study of automated evaluation and feedback for these interventions. We presented two key tasks: strategy detection and fidelity assessment, which are intended for training and evaluating AI algorithms. Furthermore, we established baseline systems for both tasks. Looking ahead, our future work will focus on refining the annotations with more detailed labels, developing a robust evaluation framework, and organizing a competition to encourage wider participation and maximize impact.

Acknowledgments. This work is supported, in part, by the National Science Foundation under Grant 2229873 (AI4ExceptionalEd) and U.S. Department of Education under Grant R305C240046 (CELaRAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

1. Akemoglu, Y., Hinton, V., Laroue, D., Jefferson, V.: A parent-implemented shared reading intervention via telepractice. Journal of Early Intervention **44**(2), 190–210 (2022)
2. Akemoğlu, Y., Laroue, D., Kudesey, C., Stahlman, M.: A module-based telepractice intervention for parents of children with developmental disabilities. Journal of Autism and Developmental Disorders **52**(12), 5177–5190 (2022)
3. Akemoglu, Y., Tomeny, K.R.: A parent-implemented shared-reading intervention to promote communication skills of preschoolers with autism spectrum disorder. Journal of Autism and Developmental Disorders **51**(8), 2974–2987 (2021)
4. Burke, M.M., Goldman, S.E.: Special education advocacy among culturally and linguistically diverse families. Journal of Research in Special Educational Needs **18**, 3–14 (2018)
5. Butler, L.K., La Valle, C., Schwartz, S., Palana, J.B., Liu, C., Peterman, N., Shen, L., Tager-Flusberg, H.: Remote natural language sampling of parents and children with autism spectrum disorder: Role of activity and language level. Frontiers in Communication **7**, 820564 (2022)
6. Cai, M., Li, M., Xiong, Z., Zhao, P., Li, E., Tang, J.: An advanced deep learning framework for video-based diagnosis of asd. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 434–444. Springer (2022)
7. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
8. Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R.M., Rozga, A., Rehg, J.M.: Detecting gaze towards eyes in natural social interactions and its use in child assessment. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(3), 1–20 (2017)
9. Dawson-Squibb, J.J., Davids, E.L., Harrison, A.J., Molony, M.A., de Vries, P.J.: Parent education and training for autism spectrum disorders: Scoping the evidence. Autism **24**(1), 7–25 (2020)
10. Duan, H., Zhai, G., Min, X., Che, Z., Fang, Y., Yang, X., Gutiérrez, J., Callet, P.L.: A dataset of eye movements for the children with autism spectrum disorder. In: Proceedings of the 10th ACM Multimedia Systems Conference. pp. 255–260 (2019)
11. Elbattah, M.: Visualization of eye-tracking scan path in autism spectrum disorder: Image dataset. In: Proceedings of the 12th International Conference on Health Informatics, Prague, Czech Republic. pp. 22–24 (2019)
12. Feng, T., Xu, A., Lahiri, R., Tager-Flusberg, H., Kim, S.H., Bishop, S., Lord, C., Narayanan, S.: Can generic llms help analyze child-adult interactions involving children with autism in clinical observation? arXiv preprint arXiv:2411.10761 (2024)
13. Holmes, W.: Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign (2019)
14. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al.: Chatgpt for good? on opportunities and challenges of large language models for education. Learning and individual differences **103**, 102274 (2023)
15. Kim, W., Seong, M., Kim, K.J., Kim, S.: Engagnition: A multi-dimensional dataset for engagement recognition of children with autism spectrum disorder. Scientific Data **11**(1), 299 (2024)
16. Kotsi, S., Handrinou, S., Iatraki, G., Soulis, S.G.: A review of artificial intelligence interventions for students with autism spectrum disorder. Disabilities **5**(1), 7 (2025)

17. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems: a meta-analytic review. Review of educational research **86**(1), 42–78 (2016)
18. Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., Rutter, M.: The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of autism and developmental disorders **30**, 205–223 (2000)
19. McWilliam, R.: Working with families of young children with special needs. Guilford Press (2010)
20. OpenAI: Chatgpt: Language model for natural language processing (2023), https://openai.com/chatgpt, accessed: 2025-02-20
21. Đorđević, M., Glumbić, N., Memisevic, H., Brojčin, B., Krstov, A.: Parent-teacher interactions, family stress, well-being, and parental depression as contributing factors to parental involvement mechanisms in education of children with autism. International Journal of Developmental Disabilities **68**(6), 838–849 (2022)
22. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
23. Serna-Aguilera, M., Nguyen, X.B., Seo, H.S., Luu, K.: A novel dataset for video-based autism classification leveraging extra-stimulatory behavior. arXiv preprint arXiv:2409.04598 (2024)
24. Sheridan, S.M., Knoche, L.L., Edwards, C.P., Bovaird, J.A., Kupzyk, K.A.: Parent engagement and school readiness: Effects of the getting ready intervention on preschool children's social–emotional competencies. Early education and development **21**(1), 125–156 (2010)
25. Soucek, T., Lokoc, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11218–11221 (2024)
26. Sukkar, H., Dunst, C.J., Kirkby, J.: Early childhood intervention: Working with families of young children with special needs. Taylor & Francis (2016)
27. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
28. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing **5**(1), 86–98 (2014)
29. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education–where are the educators? International Journal of Educational Technology in Higher Education **16**(1), 1–27 (2019)
30. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Feng, Y., Lefever, E. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-demo.49, https://aclanthology.org/2023.emnlp-demo.49/
31. Zheng, Q., Lu, X., Jin, Q., Jain, J., Meadan-Kaplansky, H., Shi, H., Xiong, J., Huang, Y.: Towards responsible use of large multi-modal ai to analyze human social behaviors. In: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing. pp. 663–665 (2024)
32. Zheng, Q., Rabbani, P., Lin, Y.R., Mansour, D., Huang, Y.: Soap. ai: A collaborative tool for documenting human behavior in videos through multimodal generative ai. In: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing. pp. 87–90 (2024)
33. Zunino, A., Morerio, P., Cavallo, A., Ansuini, C., Podda, J., Battaglia, F., Veneselli, E., Becchio, C., Murino, V.: Video gesture analysis for autism spectrum disorder detection. In: International Conference on Pattern Recognition (ICPR) (2018)