Dyadic Affect in Parent-child Multimodal Interaction: Introducing the DAMI-P2C Dataset and its Preliminary Analysis

Huili Chen\textsuperscript{1*}, Sharifa Alghowinem\textsuperscript{1}, Soo Jung Jang\textsuperscript{1}, Cynthia Breazeal\textsuperscript{1}, Hae Won Park\textsuperscript{1}

\textsuperscript{1}Personal Robots Group, MIT Media Lab, Cambridge, MA 02140

Abstract—High-quality parent-child conversational interactions are crucial for children’s social, emotional, and cognitive development. However, many children have limited exposure to these interactions at home. As increasingly accessible and scalable interventions in child development, interactive technologies, such as social robots, have great potential for facilitating parent-child interactions. However, such technology-based interventions are still underexplored, as the technologies’ limited ability to understand the social-emotional dynamics of human dyadic interactions impedes their effective delivery of timely, adaptive interventions. To advance research on resolving this roadblock, we present a “dyadic affect in multimodal interaction - parent to child” (DAMI-P2C) dataset collected during a study of 34 parent-child pairs, where parents and children (3-7 years old) engaged in reading storybooks together. In contrast to existing public datasets for social-emotional behaviors in dyadic interactions, each instance for both participants in our dataset was annotated for affect by three labelers. Additionally, the dataset contains audiovisual recordings as well as each dyad’s sociodemographic profiles, co-reading behaviors, affect labels, and body joints. We describe the dataset’s main characteristics and provide a preliminary analysis of the interrelations between sociodemographic profiles, co-reading behaviors, and affect labels. The dataset provides us with useful insights into the computing and social science fields.


1 INTRODUCTION

Sensitive, reciprocal, warm caregiver-child interactions enhance general child development\textsuperscript{1} and improve specific social, emotional and cognitive aspects of a child’s development, such as social competence\textsuperscript{2}, cooperation\textsuperscript{3} and cognition\textsuperscript{4}. Such high-quality interactions are also crucial for young children’s language development\textsuperscript{5},\textsuperscript{6},\textsuperscript{7},\textsuperscript{8}. However, a major problem faced by many children, particularly those from low socioeconomic status (SES) families, is limited exposure to socially and linguistically rich adult-child conversations at home. For instance, studies have reported that in low-SES families, caregiver-to-child conversations tend to be less frequent and of shorter duration\textsuperscript{8},\textsuperscript{9}, and have fewer open-ended questions and discussions, e.g., caregiver-child negotiations\textsuperscript{8},\textsuperscript{9},\textsuperscript{10},\textsuperscript{11}, compared with higher-SES families. Substantial disparities in vocalization and vocabulary development between children from low- and higher-SES families have also been found\textsuperscript{12} and often tend to magnify over time once children enter school\textsuperscript{8}.

It should be noted that the interest level of caregivers in their children’s education does not vary across different SES groups\textsuperscript{13}. Instead, it is the lack of access to parental education and guidance that contributes to the parental “participation gap”, namely the gap in caregivers’ active participation in their children’s education between low- and higher-SES families\textsuperscript{13},\textsuperscript{14}. As noted by Hoover et al.\textsuperscript{15}, well guided parental involvement in children’s education has a significant impact on children’s cognitive development and literacy skills, motivating interventions aimed at supporting parental involvement or facilitating parent-child interactions.

Among a variety of interventions targeting the “participation gap”, traditional home-visiting intervention programs have been shown to be effective, e.g.,\textsuperscript{16}. For example, interventionists have employed specific intervention strategies, such as coaching and modeling during triadic interactions with the parent and child together in their home, and effectively improved parents’ level of engagement in the intervention activities\textsuperscript{16}. However, in under-resourced communities, children’s access to extra-curricula support and educational resources (e.g., after-school or summer programs) is in general already limited and can be very costly\textsuperscript{17}. Availability of the traditional home-visiting interventions is even more limited, given that they require specialists to assist parent-child interactions in their homes.

Hence, there is a real need and a compelling opportunity to develop interactive technologies for such a cause, especially given their promise to deliver personalized interventions at scale in a cost-effective way. Despite their potential and recent emergence in the education domain (e.g., educational apps\textsuperscript{18},\textsuperscript{19}), still very few interactive technologies can adequately support affective and reciprocal adult-child interaction in the here and now, for example, to proactively facilitate parent-child conversations about a story\textsuperscript{20}.

To achieve effective technology-based interventions, there is urgency in understanding the dynamics of parent-child
Social, affective, and conversational interactions, in developing social-affective sensing models for such interactions, and in exploring the design space for parent-child-technology interactions. In this work, the introduced DAMI-P2C dataset and its preliminary analysis shed light on all three aspects of the research motivation.

2 RELATED WORK

2.1 Affective dynamics in parent-child interactions

Affective communication is crucial for human-human interaction and is strongly linked with learning, persuasion, and many other functions. In the context of parent-child interactions, the parent’s and child’s nonverbal behaviors are also crucial for understanding their affective dynamics. Specifically, their nonverbal behaviors are often used to measure their interaction quality, such as synchrony, engagement, and attachment. The majority of parent-child interaction quality measurement scales also include specific verbal and nonverbal behaviors indicative of dyadic reciprocity, harmonious interaction, turn-taking, and shared affect. For example, the parent-child cooperation level can be influenced by both their verbal and nonverbal social cues, such as questions/instructions, gestures, like hand movements and finger-pointing, and postures, like body enclosing and retreating. In prior work, the parent’s and child’s head or body movements were also found to be correlated to their dyadic engagement and child-parent level of synchrony.

Given the strong links between nonverbal cues and affective dynamics in parent-child interactions, advancement in automatically sensing such affective dynamics via their nonverbal cues thus has the potential to increase the efficacy of technology-based interventions in the context of parent-child interactions.

2.2 Affect sensing in interactive technologies

The ability to recognize human nonverbal cues, social signals, and emotions is crucial for interactive technologies, including virtual agents and social robots, to engage with humans in an intuitive, natural, and reciprocal way. For example, such affect-sensing ability has been identified as a fundamental robot capability necessary for higher-level competencies in human-robot interactions and contributes to a robot’s user profiling and behavior adaptation capabilities.

In the context of early childhood education, interactive technologies need to have the ability to accurately recognize the affective expressions of individual children, as children’s affective states have been shown to be crucial for their learning performance when they engage with interactive technology in educational activities. Interactive technologies with this affect-sensing ability also promote children’s learning more effectively than ones without. To build an affect-aware interactive technology, user social and affective signals can be integrated into either its behavior policy or cognitive model. Specifically, such signals can serve as human feedback on the technology’s newly executed action, when integrated into its behavior adaptation model, to personalize its interaction with individuals in real time. Similarly, the signals can be integrated into the technology’s child cognitive and skill estimation models, such as a student vocabulary acquisition model, to improve its estimation accuracy. Overall, an affect-sensing ability improves the efficacy of interactive technology to deliver timely interventions to individual users and enhance their interaction experience.

The interaction paradigm of most affect-aware interactive technologies only involves a single person and a single learning technology, not including scenarios with two users and one technology. When interacting with a single person, a technology only needs to recognize one person’s social and affective cues in response to either the interaction task or the technology. Conversely, when engaging in a dyadic human interaction, the interactive technology needs to have an additional ability to understand the affective and social dynamics between the two users. However, the majority of current affect-sensing models, especially commercial affect extraction tools, are only applicable in single-person contexts.

Very few recent works have started to develop dyadic affect-sensing models. For example, Chen and colleagues recently used end-to-end deep learning methods augmented with attention mechanisms to recognize each person’s affective expression in an audio stream containing the utterances of two speakers. However, this work is limited in that the models predict each person’s affect using only an audio modality. People’s affect can be indeed predicted using a variety of modalities (i.e., audio, facial expressions, HR, EDA, temperature, and ECG), and leveraging multiple modalities for affect recognition can increase the robustness and applicability of models to challenging real-world cases. Building multimodal two-person affect-sensing models would further unlock the potential of interactive learning technologies to improve the social-emotional interactions between two people, including parent-child interactions.

Overall, there has been very little research into multiperson affect-sensing, largely limiting the development of affect-aware interactive technologies suitable for dyadic human interactions. One of the main obstacles for developing dyadic affect sensing is the lack of multimodal dyad datasets with proper affect annotations. Hence, our DAMI-P2C work contributes to the research on multimodal two-person affect sensing by providing a dataset with affect annotations, multiple modalities, and sociodemographic profiles of human dyads.

2.3 Available multimodal datasets on dyadic human-human social-emotional interactions

Datasets of dyadic human-human social interactions are crucial for understanding the way people engage with each other and for designing social technologies that can interact with people. Currently available datasets that target human-human social-emotional communication can be classified based on the target populations’ age groups: adult-adult interaction (AAI), child-child interaction (CCI), and adult-child interaction (ACI).

As the most widely examined interaction type, AAI has been extensively studied in a variety of corpora (
Though these prior adult-child interaction datasets offer rich perceived affect. In contrast, our DAMI-P2C dataset uniquely addresses these limitations above, thereby helping advance research on multi-interlocutor affect-sensing and technology-based interventions on parent-child interactions.

3 DAMI-P2C DATASET OVERVIEW

Automatic dyadic affect recognition can potentially enable interactive technologies to provide right-in-time, adaptive interventions to human dyads. However, this area of research is still underexplored partly due to the lack of dyadic human interaction datasets annotated for this type of modeling task. Hence, we present a “dyadic affect in multimodal interaction - parent to child” (DAMI-P2C) dataset aimed at capturing natural story-reading interactions between a parent and their child in a lab setting. It is the first dyadic dataset that focuses on affective parent-child interactions and provides both interlocutors’ affective states labeled by three independent labelers in short audiovisual segments, along with their sociodemographic profiles, verbal behaviors, and 2D and 3D body joints. The dataset consists of five major categories of content necessary to understand the social-emotional behaviors and affective states of parent-child dyads in the co-reading context, as presented below (see Table 1). In this dataset, we focused on the parent-child co-reading interaction activity, a practice positively associated with both children’s later reading and language outcomes [52] and their interest and enjoyment in reading later in childhood [53].

Practice child-parent dialogic reading has also shown crucial for children’s emergent literacy skills [54], [55], [56], [57]. In addition, a few learning technologies have been designed to enhance parent-child co-reading, e.g., e-book apps [58], augmented reality (AR) picture books [59], and a story discussion suggestion platform [60]. These works all together indicate that the parent-child co-reading activity has great potential for developing technology-based interventions to promote reciprocal conversational parent-child interaction. Hence, understanding parent-child affective dynamics and developing affect sensing models in this interaction activity becomes important.

Audio and video: To capture the interaction from different views, seven cameras in total were set up for audio-video recordings along with an MXL AC404 USB conference microphone. Three cameras captured the parent-child dyad’s front view and parent’s and child’s centered views, while the additional two cameras captured the dyad from a bird’s-eye view (see Fig. 1). Lastly, two wide-angled GoPro cameras were used to capture first-person views attached to the parent’s forehead and the interaction view attached to the tablet, respectively. All video and audio recordings were approximately time-synchronized. See Section 4.2.1 for details.

Sociodemographic profiles: To shed light on the potential causes of variations across parent-child social-affective interaction styles, their demographic and socio-emotional parameters were collected for each parent-child dyad through self-reported questionnaire responses. The captured parameters included the participant’s gender, age, English proficiency, ethnicity, social-economic status, and home literacy environment. In addition, the child’s temperament, parent’s parenting style, and parenting stress were captured. See Section 4.2.2 for details.

Reading behavior features: Parent-child co-reading was transcribed and annotated for each dyad to reveal their story...
reading styles. A list of dyadic reading behavior features is extracted, including conversation duration and frequency, turn-taking rates in reading and conversations, individual speaker’s speech rate, and the ratio of conversation to reading, as well as parent-child relative reading duration, conversation duration, and conversation initiations. Both raw transcripts and extracted reading behavior features are provided in the dataset. See Section 4.2.3 for details.

Affect annotations: Audio-video recordings were annotated for a wide range of behavior-based affective states to help reveal the social-emotional dynamics between the parent and the child during co-reading. The annotated affective states included valence and arousal of both the adult and child as well as engagement and coordinated engagement of the child. See Section 5 for details.

Person identification and body tracking: From the front camera view, each person in the videos was detected, tracked, and named in every frame along with their extracted 2D and 3D body joints, head points, and facial landmarks. See Section 6 for details.

4 Dataset Design

This section describes the study procedures of our data collection, as well as measurements with respect to audiovisual recordings, dyadic sociodemographic profiles, and reading behaviors, along with some initial behavioral statistics.

4.1 Data collection protocol

4.1.1 Study participants

We recruited 34 families with children between the ages of 3 and 7 years old from the greater Boston area with their full consent for two-session data collection, which occurred in our lab space. The dyad was always a pair of one parent and one child. Two parents did not report their children’s age, and four parents did not report their own. Three families withdrew from the data collection after the first session for reasons not related to the protocol. In sum, the DAMI-P2C dataset consists of two sessions of 31 families and one session of three families. The data collection was conducted in the Spring semester of 2019, between February and May. The average ages of parents and children were 39.70 ± 5.47 and 5.49 ± 1.37, respectively. The gender identity, age range, and language proficiency of the participant dyads are summarized in Table 1.

4.1.2 Study procedure

The collection protocol consisted of two 45-minute in-lab sessions where a parent and their child read stories together for 20-30 minutes, and the parent filled out surveys for the remaining 25 minutes. Two in-lab sessions were conducted within two weeks. Families that completed both sessions were given $75 as compensation. In the first session, the experimenter read the written overview of the study to the parent and child, including the study’s objective, procedures, and instructions, to ensure its consistency across individual parent-child pairs.

4.1.3 Story-reading activity

During the story-reading interaction, the parent and child sat next to each other, as shown in Fig. 1. A digitized version of our storybook corpus on a touchscreen tablet was used for the sessions (Fig. 2). The parent and child were allowed to either hold the tablet in their hands or put the tablet on the table when reading stories, but they were instructed to stay in the story station area during the reading activity. Before the story-reading activity started, the experimenter guided the dyad on how to use the tablet and the storybook app.

The storybook corpus consists of 30 storybooks recommended by early childhood education experts and teachers.
Fig. 1: Audiovisual Recordings of Parent-Child Story-reading Interactions. The DAMI-P2C dataset includes recordings of 34 families reading stories for two in-lab sessions. Seven approximately time-synchronized cameras captured the front view and bird’s-eye view of the dyad along with a child-centered view, a parent-centered view, a parent-first-person view and a tablet view. Fig. (a) displays a story-reading station setup including six different camera angles (seven cameras), a high-quality microphone, a story-reading table, and a story-reading tablet with storybooks. Fig. (b) displays an example scene of a parent-child story-time interaction; annotators viewed audiovisual recordings similar to this scene, except showing the faces, when coding parent’s and child’s affect attributes. Fig. (b)-(g) show the video recordings concurrently captured from different camera views.

Each story lasts between 3 and 15 minutes. Stories shorter than five minutes were categorized as short stories and the rest as long stories. During the story-reading session, the parent and child could select any book they have wanted to read from the corpus. Given the diversity of our participants’ backgrounds, e.g., age, culture, English proficiency, as shown in Fig. 2 and Fig. 3, selecting a fixed small set of stories for all dyads to read would not be suitable. Any of their potential negative affects or uncooperative behaviors arisen from disliking the assigned stories may contaminate the collected affective data aimed at revealing parent-child dynamics in their everyday reading practice. Instead, allowing them to pick their preferred books would allow the collected dataset to more accurately reflect their natural and typical story-reading behaviors. To further ensure the dataset’s naturalistic quality, we also encouraged the dyads to read stories together in a way they would normally do at home, having conversations about the stories and making the activity fun and interactive.

4.2 Measurements

4.2.1 Audiovisual data recording and synchronization

During the in-lab parent-child co-reading sessions, audiovisual recordings were captured using one microphone and seven cameras installed in the story-reading station (Fig. 1). Five Logitech c930e cameras were fixed in portable walls and used to capture four different angles of the dyadic interaction (i.e., frontal view, bird’s-eye view, parent-centered view, and child-centered view). In addition to the fixed cameras, we used two wide-angle cameras (GoPro HERO Session): one attached to the parent’s forehead to capture their first-person view, and a second attached to the tablet to capture a close-up view of the parent-child interaction. Audio recordings were captured using a microphone (MXL AC404 USB conference microphone, three-capsule design with 25’ audio pick up in a

<table>
<thead>
<tr>
<th>Gender</th>
<th>Parent (age)</th>
<th>Child (age)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>25 (38.78 ± 4.70)</td>
<td>13 (5.20 ± 1.96)</td>
</tr>
<tr>
<td>Male</td>
<td>9 (42.25 ± 6.90)</td>
<td>21 (5.65 ± 0.96)</td>
</tr>
<tr>
<td>N/B</td>
<td>28 (39.54 ± 5.19)</td>
<td>29 (5.58 ± 1.46)</td>
</tr>
<tr>
<td>ELL</td>
<td>6 (40.33 ± 7.15)</td>
<td>5 (5.00 ± 0.64)</td>
</tr>
</tbody>
</table>

Fig. 2: A digitized version of the storybook corpus on the touchscreen tablet. The books were divided into two categories based on story length.

TABLE 2: Gender identity, age range, and language proficiency of the participant families in the dyadic dataset. English language learner is denoted as ELL. Native/bilingual speaker is denoted as N/B.
180° arc). The video recordings had a resolution of 720x1280 at 30 Hz, except for the additional video recording from the bird’s-eye view that had a resolution of 480x640 at 30 Hz, and the audio recordings were 16-bit at 44.1 kHz. The audio recordings were sent to a professional transcription service to obtain textual annotations of the recorded speech and timestamps for conversational turn-takes in the format of comma-separated values (CSV) files.

Given the complex recording setting with seven cameras and a microphone, not all devices were synchronized at the time of recording. Therefore, we approximated the synchronization of the footage off-line, using the timestamps printed in the top-left corner of each frame of the Logitech cameras and the file timestamps of the GoPro cameras. In each session, we used one of the bird’s-eye views as the main video, where each remaining video was either filled with blank frames or the extra frames were removed from the beginning and the end of the video to be aligned with the main video and to ensure all videos had the same duration as the main one. However, since the timestamps did not include milliseconds, and since the frame rate is not equal at every time slice, the synchronization is not perfectly aligned, but its rough approximation is around 1 second off. People interested in a perfect alignment could match the timestamps of every single frame from each camera and use the number of frames between every two seconds to count the frame rate in each time slice. The meta-data used for this approximation of synchrony is included in the dataset (in case reverse actions are required). A summary of the available video files are listed in Table 3 and specific details of missing files are provided in the dataset documentation.

**TABLE 3: A summary of the available video recordings.** The total number of conducted sessions is 65, where 31 parent-child dyads finished both sessions and 3 parent-child dyads only finished the first session.

<table>
<thead>
<tr>
<th>Camera View</th>
<th>Total Sessions</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bird’s-eye 1 (Fig. 1a)</td>
<td>65</td>
<td>34</td>
<td>31</td>
</tr>
<tr>
<td>Bird’s-eye 2 (Fig. 1b)</td>
<td>61</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Front (Fig. 1c)</td>
<td>59</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>Parent-centred (Fig. 1e)</td>
<td>64</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>Child-centred (Fig. 1f)</td>
<td>65</td>
<td>34</td>
<td>31</td>
</tr>
<tr>
<td>Parent first-person (Fig. 1g)</td>
<td>53</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td>Tablet first-person (Fig. 1h)</td>
<td>49</td>
<td>25</td>
<td>24</td>
</tr>
</tbody>
</table>

**4.2.2 The sociodemographic profiles of parent-child dyads**

The sociodemographic profiles of parent-child dyads were collected using questionnaires completed by the parents. In total, we collected participant families’ demographic information, home literacy environment [61], parenting style [62], parenting stress [63], and child’s temperament and behavior [64]. The four families who only completed the first-in-lab session did not fill out the questionnaire. In total, the profile data consists of five sociodemographic categories - Child’s Behavior Questionnaire (CBQ), Parenting Relationship Questionnaire (PRQ), Parenting Stress Index (PSI), Demographic Profile (DEMO), and Home Literacy Environment (HLE), resulting in 17 sub-scale sociodemographic features and three aggregated total score features. The sociodemographic data of 30 families are reported in Table 4 and Fig. 3 and used for the feature analysis in Section 7.

**TABLE 4: Summary of parent-child dyads’ sociodemographic profiles.** The profile data consist of five profile categories (i.e., CBQ, PRQ, PSI, DEMO, and HLE) with 17 sub-scale sociodemographic features and 3 aggregated total score features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean ± SD</th>
<th>Min - Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBQ</td>
<td>4.37 ± 0.50</td>
<td>3.61 - 5.42</td>
</tr>
<tr>
<td>Surgency</td>
<td>4.39 ± 0.58</td>
<td>3.08 - 5.50</td>
</tr>
<tr>
<td>Negative Affect</td>
<td>4.03 ± 0.63</td>
<td>2.77 - 5.07</td>
</tr>
<tr>
<td>Effortful Control</td>
<td>5.31 ± 0.81</td>
<td>3.25 - 6.75</td>
</tr>
<tr>
<td>PRQ</td>
<td>74.00 ± 14.52</td>
<td>43 - 99</td>
</tr>
<tr>
<td>Attachment</td>
<td>31.67 ± 4.98</td>
<td>24 - 41</td>
</tr>
<tr>
<td>Discipline Practices</td>
<td>12.43 ± 2.94</td>
<td>7 - 20</td>
</tr>
<tr>
<td>Involvement</td>
<td>18.37 ± 3.81</td>
<td>10 - 26</td>
</tr>
<tr>
<td>Parenting Confidence</td>
<td>26.43 ± 5.11</td>
<td>16 - 36</td>
</tr>
<tr>
<td>Relational Frustration</td>
<td>14.90 ± 4.69</td>
<td>8 - 32</td>
</tr>
<tr>
<td>PSI</td>
<td>73.20 ± 19.76</td>
<td>45 - 128</td>
</tr>
<tr>
<td>Distress</td>
<td>22.90 ± 7.12</td>
<td>13 - 42</td>
</tr>
<tr>
<td>Dyad Dysfunction</td>
<td>22.77 ± 6.87</td>
<td>15 - 40</td>
</tr>
<tr>
<td>Difficult Child</td>
<td>27.53 ± 7.62</td>
<td>17 - 46</td>
</tr>
<tr>
<td>DEMO</td>
<td>66.70 ± 11.89</td>
<td>43 - 83</td>
</tr>
<tr>
<td>Child’s Age (Months)</td>
<td>2.72 ± 0.88</td>
<td>1 - 4</td>
</tr>
<tr>
<td>Parent’s Education</td>
<td>0.53</td>
<td>0 - 1</td>
</tr>
<tr>
<td>English at Home</td>
<td>1.27 ± 0.78</td>
<td>0 - 2</td>
</tr>
<tr>
<td>Parent’s Years in USA</td>
<td>4.23 ± 2.05</td>
<td>0 - 6</td>
</tr>
<tr>
<td>HLE</td>
<td>0.65 ± 0.14</td>
<td>0.44 - 0.93</td>
</tr>
</tbody>
</table>

**Child’s Behavior Questionnaire (CBQ):** The CBQ questionnaire [64] assesses a child’s temperament in early to middle childhood. We used a 36-question version with three main sub-scales: Surgency, Negative Affect, and Effortful Control. Surgency averages the child’s activity level, high-intensity pleasure, impulsivity, and negative shyness. Negative affect averages the child’s anger, discomfort, fear, sadness, and negative soothability. Effortful Control averages the child’s attention focusing, inhibitory control, low-intensity pleasure, and perceptual sensitivity. Each response is recorded using a 7-item Likert scale, and the score of each sub-scale is an average of corresponding Likert responses. The total CBQ score is an average of all Likert responses, with a higher score representing a higher level of negative temperament. As shown in Fig. 3e, the CBQ distributions were not heavily skewed toward either extreme, indicating that the children in our study tended to have moderate temperament.

**Parenting Relationship Questionnaire (PRQ):** The PRQ questionnaire [62] measures a caregiver’s parenting style. It is comprised of 60 questions, resulting in five sub-scales: Attachment, Discipline Practices, Involvement, Parenting Confidence, and Relational Frustration. Each response is on a Likert scale between 0 and 3 and the score of each sub-scale is an average of corresponding Likert responses. The total PRQ score is a modified sum of PRQ sub-scale scores with the Relational Frustration score inverted; a higher PRQ total score means higher parental attachment, discipline, involvement, and confidence, and lower frustration. As shown in Fig. 3f, the four PRQ sub-scale features shared distinct distribution patterns with Parenting Confidence having the most widely spread distribution (Fig. 3f), suggesting that the sub-scale features capture distinct aspects of parent-child relationships in our dataset.

**Parenting Stress Index (PSI):** The PSI questionnaire [63]...
measures a parent’s level of parenting stress. Comprised of 36 questions, it represents multiple parental stress indicators, namely Parental Distress, Dyad Dysfunction, and Difficult Child. Each response is a 5-item Likert scale, and each subscale score is calculated by summing its individual responses. A higher PSI score means a higher parenting stress level. As shown in Fig. 3, the distributions are more skewed towards the low parenting stress side.

**Demographic profile (DEMO):** Parent-child dyads’ demographics and cultural backgrounds were measured with five parameters relevant to the shared reading context [65]. Whether English is used at home, i.e., English at Home, was recorded as binary responses with 1 for English used at home. The Child’s Age was calculated from the children’s birthdays as integer months. Parent’s Years in the USA was chosen from the 7-item Likert scale options, ranging from “0 to 1 year” (score 0) to “more than 30 years” (score 6). The Parent’s English Level was chosen from the 3-item Likert scale options: “second language learner” (score 0), “bilingual” (score 1), and “native” (score 2). Parent’s Education was chosen from the 5-item Likert scale options, ranging from “high school or less” (score 0) to “doctoral-level degree” (score 4). As shown in the distributions in Fig. 3, children varied in age and parents had a moderately wide range of education. Parents’ English proficiency and years in the USA were more skewed towards being native or having resided long in the USA. Lastly, approximately half of the families used English at home.

**Home Literacy Environment (HLE):** Parent-child dyads’ HLE was measured using responses to six questions: number of children’s books in the home, age of the child when first read to, amount of time spent on reading with the child at home, amount of time spent by the child reading alone, frequency of family members reading to the child, and frequency of family members teaching child alphabet [61]. Responses to each question were measured on a 5-item Likert scale and converted to the percentage of the maximum possible score, and the total HLE score was calculated by averaging the responses to the six questions. A higher HLE score represents a better home literacy environment. The HLE distribution in Fig. 3 shows that participants’ home literacy environment varied in our dataset.

### 4.2.3 Reading behavior feature extraction

Dyadic reading behavior features were extracted from the text transcripts of parent-child speech during the co-reading activity. As suggested by Whitehurst *et al.* [66], good parent-child shared reading practices include the following three dimensions: (1) dialogic reading (i.e., discussions between parent and child in parallel with story reading), (2) high child participation (i.e., high portions of the child reading and speaking) and (3) active parent-child turn-taking (i.e., frequent switching of reading or speaking turns between child and parent).

The above three reading behavior dimensions were contextualized in our dataset to quantify parent-child shared reading quality in the following way. We first distinguished between parent-child story reading and conversing. Story reading is defined as the parent or child reading a storybook verbatim, while conversing refers to the parent-child conversations about the storybook they are reading (i.e., dialogic reading). To extract features related to dialogic reading and

2. Notice that the scores of Parent’s English Level and Child’s Age presented in Table 4 differ from those in Table 2, as the data of only 30 families were used as the sociodemographic features to keep consistency across all profile categories for the feature analysis in Section 7.
child’s participation, we compared the transcript of parent-child speech against the storybook script and measured portions of the transcript that are parent-child reading or conversing. In total, nine dyadic reading behavior features were extracted to characterize the three reading behavior dimensions in our context, and parent’s and child’s speech rates were also added as two additional features to reflect the pace of the co-reading activity (Table 5).

5 AFFECT AND ENGAGEMENT ANNOTATIONS

5.1 Valence and arousal

An individual’s affective expressions can be recognized using different affect scales. As one of the most extensively used scales, the valence–arousal scale encompasses all possible affective states and their variations [67], with the valence capturing a user’s level of pleasure and arousal, revealing a user’s excitement level [68]. Compared with categorical emotional models that capture emotions such as happiness, anger and sadness, the valence-arousal scale also represents experimental and clinical findings more accurately [69]. In addition, the perceived valence and arousal of an individual have also been used as two components of affective engagement in technology-assisted interactions before, such as in autism therapy [70]. Hence, the valence-arousal scale was selected for affect annotations of our dataset.

In our dataset, we only measured the outward behavioral characteristics of valence and arousal, e.g., facial expressions [71], speech [36], and body postures [72], although arousal and other facets of affect can also be measured from inward cues, e.g., physiological signals [73]. Both valence and arousal of the child and parent during co-reading were annotated using their verbal and nonverbal behaviors observed in the audiovisual recordings.

5.2 Engagement

Engagement, that is, a process where multiple agents establish, maintain, and agreeably end their perceived connection during a joint interaction [74], has been widely used to measure the quality of human-human interactions. In the context of interactive technologies, user engagement is also often measured to make inferences about human-technology interaction quality (e.g., the human-robot engagement model proposed by Salam and Chetouani [75]). In addition, accurately detecting the state of user engagement would enable interactive technologies to deliver timely interventions [76]. For these purposes, user engagement is also included as a user affect feature in DAMI-P2C.

To capture the parent-child engagement quality, we chose the Joint Engagement Rating Inventory (JERI) [77], as it quantitates and qualifies the caregiver-child interaction during a joint activity where verbal and nonverbal behaviors related to engagement are observed and rated. JERI comprises more than ten scales capturing diverse aspects of parent-child interaction, and the two engagement scales, i.e., Child Unengaged and Child Coordinated Engagement, were selected for our annotation. Child Unengaged captures the child’s overall engagement with both the parent and the activity. When the child is not actively attentive to both the parent and the reading activity (book/tablet), the child is considered unengaged. In our dataset, we inverted the score of Child Unengaged as a measure of Child’s Engagement (CE). According to JERI, Child Coordinated Engagement (CCE) involves the child’s engagement with the parent instead of their engagement with the activity. The child’s CCE will be rated low if the child is engaging in story listening or reading without attending to the parent and acknowledging their presence.

5.3 Annotation

We recruited five trained annotators with a psychology or education background to independently annotate the audiovisual recordings of the families’ co-reading interactions. In total, six labels were each annotated separately by three annotators. While watching the recordings, the coders gave ratings every five nonoverlapping seconds on a five-point ordinal scale [-2,2], with 2 corresponding to cases when the target person showed clear signs of high arousal/positive valence/high engagement/high coordinated engagement and -2 when the person showed clear signs of low arousal/negative valence/low engagement/low coordinated engagement.

A five-second window was selected as the fragment interval of target audiovisual recordings for the annotation to produce continuous quality scales, a threshold consistent with prior work on affect detection [78]. When annotating the recordings, the annotators were instructed to judge whether a given fragment contained a story-related dyadic interaction and filter out those that did not. In total, 16,593 five-second fragments were annotated, with 488.03 ± 123.25 fragments from each family on average, as shown in Table 3.

The agreement of the three annotators was measured using the intra-class correlation (ICC) [29] type (3,1) for average fixed raters. The ICC is commonly used in behavioral sciences to assess annotators’ agreement, and the score ranges from 0-100%. The average ICC and its confidence interval among the annotators for each of our labels are presented in Table 4. According to the commonly-cited cutoffs for qualitative inter-rater reliability (IRR) based on ICC values [60], [61], IRR is good if $0.60 \leq ICC < 0.75$, and IRR is excellent if $0.75 \leq ICC \leq 1.0$. Given this evaluation criteria, the annotations for the six affect attributes achieve either good or excellent quality. After the recordings were coded independently by the annotators, we took the average of a scale’s ratings from the three annotators for each recording fragment as its final score of the scale.

The label distribution for each target affect attribute after the individual ratings were averaged, as depicted in Fig. 4, and the pairwise Spearman correlation coefficients between the six attributes are reported in Table 7. The correlation results showed that speaker valence and arousal are moderately positively correlated with each other for both parent and child. Additionally, the parent’s and child’s arousal states are negatively correlated with each other ($\rho = -0.27$). The latter may be explained by the conversational interactions, in which the parent and the child took turns being speaker and listener. The correlation results also showed that child’s coordinated engagement has moderately strong positive

3. Coding manuals for the six affect attributes are on the DAMI-P2C dataset website.
Within an interaction, their affective behaviors might vary (see Fig. 1b) in the following steps: Child’s Coordinated Engagement (CCE), Parent’s Valence (PV), Child’s Engagement (CE), and Parent’s Arousal (PA).

TABLE 5: Definitions and summary statistics of parent-child reading behavior features. The parent-child conversation is denoted as Conv., and a conversation session is defined as the period after the dyad switches from reading to conversing and before they switch from conversing back to reading.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Dyadic Reading Feature</th>
<th>Description</th>
<th>Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialog</td>
<td>Conv.: Read Ratio of Parent</td>
<td>The ratio of parent’s time spent on conversations to their time spent on reading</td>
<td>0.96 ± 1.34</td>
</tr>
<tr>
<td></td>
<td>Conv.: Read Ratio of Child</td>
<td>The ratio of child’s time spent on conversations to their time spent on reading</td>
<td>7.43 ± 15.53</td>
</tr>
<tr>
<td></td>
<td>Conversation Duration</td>
<td>Average time spent on each conversation session in seconds</td>
<td>11.39 ± 8.63</td>
</tr>
<tr>
<td></td>
<td>Conversation Frequency</td>
<td>Total # of conversations divided by total story reading length</td>
<td>0.04 ± 0.02</td>
</tr>
<tr>
<td>Child’s</td>
<td>Child’s Portion in Reading</td>
<td>The portion of child’s reading time relative to total reading time</td>
<td>0.19 ± 0.25</td>
</tr>
<tr>
<td>Participation</td>
<td>Child’s Portion in Initiation</td>
<td>The portion of child-initiated conversations relative to total conversation initiations</td>
<td>0.29 ± 0.14</td>
</tr>
<tr>
<td>Active Turn</td>
<td>Turn-Taking Rate in Reading</td>
<td>Total # of turns in story reading divided by total story reading length</td>
<td>0.04 ± 0.04</td>
</tr>
<tr>
<td>Taking</td>
<td>Turn-Taking Rate in Conv.</td>
<td>Total # of conversational turns divided by total conversation length</td>
<td>0.12 ± 0.05</td>
</tr>
<tr>
<td>Speech</td>
<td>Parent’s Speech Rate</td>
<td>Average parent’s uttered word count per second</td>
<td>2.07 ± 0.61</td>
</tr>
<tr>
<td>Rate</td>
<td>Child’s Speech Rate</td>
<td>Average child’s uttered word count per second</td>
<td>1.38 ± 0.47</td>
</tr>
</tbody>
</table>

TABLE 6: Statistics of the number of affect and engagement annotations collected across families. The first, second, and third quartiles are denoted as Q1, Q2, and Q3, respectively.

<table>
<thead>
<tr>
<th>Mean ± SD</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>488.03±123.23</td>
<td>235</td>
<td>724</td>
<td>438.25</td>
<td>482.00</td>
<td>576.75</td>
</tr>
</tbody>
</table>

TABLE 7: Pairwise Pearson correlation coefficients between Child’s Arousal (CA), Parent’s Arousal (PA), Child’s Valence (CV), Parent’s Valence (PV), Child’s Engagement (CE), and Child’s Coordinated Engagement (CCE).

<table>
<thead>
<tr>
<th></th>
<th>CA</th>
<th>PA</th>
<th>CV</th>
<th>PV</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>-0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>0.52</td>
<td>-0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PV</td>
<td>-0.19</td>
<td>0.43</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CE</td>
<td>0.14</td>
<td>-0.21</td>
<td>0.14</td>
<td>-0.10</td>
<td></td>
</tr>
<tr>
<td>CCE</td>
<td>0.22</td>
<td>0.00</td>
<td>0.28</td>
<td>0.07</td>
<td>0.32</td>
</tr>
</tbody>
</table>

correlations with the child’s engagement ($\rho = 0.32$), child’s arousal ($\rho = 0.22$), and child’s valence ($\rho = 0.28$). As shown in Fig. 4, the annotations for each affect label are centered around certain values and not balanced. Given the natural story-reading interaction context, the affective behaviors of the participants were not explicitly and purposefully elicited and thereby likely to be centered around some baselines. Within an interaction, their affective behaviors might vary or deviate from the baselines when having conversations or encountering emotion-eliciting stimuli from reading.

6 PERSON IDENTIFICATION AND BODY TRACKING

In order to facilitate and benchmark future automatic affect analyses using the DAMI-P2C dataset, we provide identified bounding boxes for each person in the scene as well as their 2D and 3D body joints. For this purpose, we developed a pipeline to identify and continuously track the parent’s and child’s bodies in the main video of the front-camera view (see Fig. 11) in the following steps:

(1) Detecting human bodies to analyze the activities performed by each person in the scene. Common object detection models, e.g., YOLO v3 [82] and Detectron v2 [83], vary in their accuracy and speed. In our context, YOLO v3 balances both accuracy and speed while Detectron v2 achieves a high object detection accuracy at a lower frame rate. Given the relatively low accuracy of YOLO v3 in detecting child bodies, we sampled ~50k frames, in which YOLO failed to detect both the parent’s and child’s bodies, to be further annotated by Detectron v2 [83] for the missing bodies. The annotations returned by Detectron v2 along with the frames were then used to retrain the YOLO model customized for detecting child’s bodies in our dataset.

(2) Continuously tracking the detected bodies using a hy-

TABLE 8: The average score of annotation and the average score and confidence interval (CI) of intra-class correlation (ICC) for each affect attribute.

<table>
<thead>
<tr>
<th>Label</th>
<th>Mean ± SD</th>
<th>ICC score</th>
<th>ICC CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>1.42±0.88</td>
<td>0.84</td>
<td>[0.84, 0.85]</td>
</tr>
<tr>
<td>PA</td>
<td>2.07±0.53</td>
<td>0.83</td>
<td>[0.82, 0.83]</td>
</tr>
<tr>
<td>CV</td>
<td>2.14±0.31</td>
<td>0.64</td>
<td>[0.63, 0.65]</td>
</tr>
<tr>
<td>PV</td>
<td>2.57±0.37</td>
<td>0.61</td>
<td>[0.60, 0.62]</td>
</tr>
<tr>
<td>CE</td>
<td>3.10±0.68</td>
<td>0.81</td>
<td>[0.80, 0.81]</td>
</tr>
<tr>
<td>CCE</td>
<td>2.52±0.64</td>
<td>0.80</td>
<td>[0.80, 0.81]</td>
</tr>
</tbody>
</table>

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3178689, IEEE Transactions on Affective Computing

(4) Extracting 2D body joints to calculate and analyze the geometrical and temporal movement features of individual bodies and in relation to each other. For this purpose, we used AlphaPose [87] to extract 17 body joints in the 2D space, which has higher accuracy in comparison with its counterparts (e.g., OpenPose [88]).

(5) Estimating the 3D body joints from the 2D joints using VideoPose3D [89], a state-of-the-art model that utilizes the temporal aspects of 2D joints to predict the joints in the 3D space. It is important to notice that the 3D triangulation algorithm only estimates body pose given the 2D points without depth estimation (see Fig. 5).

The identified bounding boxes and both the 2D and 3D body joints are provided in the DAMI-P2C dataset for each frame in the front-view video (Fig. 5). They create more opportunities for analyzing the nonverbal behaviors of individuals and nonverbal interpersonal dynamics in dyadic parent-child interactions as well as their relations to the affect, engagement, sociodemographic profiles, and reading behaviors of the dyads provided in the dataset. These annotations are organized by family ID and session number, where per-frame annotations (i.e., compressed Numpy files) are zipped. For each person in each frame, the data contains the bounding box, its ID, its name (i.e., parent, child, others), the 2D and 3D joints (in the dataset technical report, we provide a code snippet on how to deal with the frame data). Moreover, we provide the camera matrices (intrinsic and extrinsic) along with calibration images from each camera in order to estimate the camera matrices, which could be particularly beneficial for 3D analysis.

7 FEATURE ANALYSIS AND RESULTS

To highlight potential candidate feature associations for a future in-depth investigation, we conducted exploratory correlation analyses between the sociodemographic features, between the sociodemographic and reading features, and between the sociodemographic and affect features by calculating pairwise Spearman correlation coefficients. The 17 sociodemographic features, 11 reading features, and 6 affect features used in the correlation analyses are presented in Table 4, Table 5, and Table 8 respectively. For the analyses involving the sociodemographic features, the aggregated total score for each sociodemographic category was excluded, and only their individual sub-scale features were included in the analyses to avoid redundancy and repetitiveness. Since the degree of significance depends upon the sample...
size and the effect size, the desired correlation coefficient cutoff was calculated to determine whether a coefficient differs from zero given our sample size ($n = 30$), using the standard formula for sample size estimation. In our case, the desired coefficient cutoff is $|\rho| \geq 0.49$, and hence, any correlations with $|\rho| < 0.49$ would not be considered significant in our correlation analysis. $p$-value is not used because the correlation analyses only aim to show potential candidate feature correlations that may be worth examining further. The analysis is exploratory rather than confirmatory, only to reveal the descriptive statistics of the dataset. In this regard, the post hoc $p$-value correction for multiple testing is also not applicable and thereby not performed in the analysis.

### 7.1 Pairwise correlations between sociodemographic features

Pairwise correlations were calculated between the sub-scale features of five different sociodemographic categories, i.e., CBQ, PRQ, PSI, DEMO, and HLE. Correlations within the same sociodemographic categories were excluded to reduce redundancy and repetitiveness. The results show that 7 of the 110 pairwise correlations were significant ($|\rho| \geq 0.49$).

As shown in Table 9, all the top five correlated feature pairs with the highest strength scores are from PRQ and PSI, indicating that these two profile categories are more closely linked with each other than the others. For example, PRQ Frustration is most strongly correlated with PSI Parental Distress ($\rho = 0.80$), probably because these two features reflect very similar underlying psychological constructs.

**TABLE 9: Significant correlations between the sociodemographic sub-scale features ($|\rho| \geq 0.49$).**

<table>
<thead>
<tr>
<th>Sub-scale Feature</th>
<th>Sub-scale Feature</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRQ Frustration</td>
<td>PSI Parental Distress</td>
<td>0.80</td>
</tr>
<tr>
<td>PRQ Frustration</td>
<td>PSI Difficult Child</td>
<td>0.69</td>
</tr>
<tr>
<td>PRQ Confidence</td>
<td>PSI Parental Distress</td>
<td>-0.67</td>
</tr>
<tr>
<td>PRQ Frustration</td>
<td>PSI Dyad Dysfunction</td>
<td>0.60</td>
</tr>
<tr>
<td>PRQ Attachment</td>
<td>PSI Dyad Dysfunction</td>
<td>-0.54</td>
</tr>
<tr>
<td>PRQ Parenting Confidence</td>
<td>DEMO Age (Months)</td>
<td>-0.50</td>
</tr>
<tr>
<td>PRQ Attachment</td>
<td>DEMO Child’s Age (Months)</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

### 7.2 Pairwise correlations between sociodemographic and reading features

The results of the pairwise correlation analysis between the sociodemographic and reading behavior features show that 5 of the 187 correlations are significant ($|\rho| \geq 0.49$). As shown in the correlation list (Table 10), parent’s speech rate during story reading is positively influenced by their use, exposure, and proficiency of English (DEMO Parent’s English Level: $\rho = 0.78$; DEMO Parent’s Years in USA: $\rho = 0.58$; DEMO English at Home: $\rho = 0.50$), reflecting that parents who are more comfortable with English are more likely to read and speak faster. Furthermore, the frequency of parent-child conversations is negatively correlated to child’s age ($\rho = -0.54$), suggesting that shared reading sessions with younger children yield more frequent conversations. Lastly, PRQ Relational Frustration is negatively correlated with parent-child turn-taking rate of reading. This correlation result indicates that a parent’s frustration with their parent-child relationship could be revealed in their fewer back-and-forth interactions with their child in story-reading. Among the five correlations, four are associated with the DEMO category, indicating that sociodemographic parameters, in particular parent’s language-related ones, play crucial roles in influencing parent-child reading behaviors.

**TABLE 10: Significant correlations between sociodemographic and reading behavior features ($|\rho| \geq 0.49$).**

<table>
<thead>
<tr>
<th>Reading Feature</th>
<th>Sociodemographic Feature</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent’s Speech Rate</td>
<td>DEMO Parent’s English</td>
<td>0.78</td>
</tr>
<tr>
<td>Parent’s Speech Rate</td>
<td>DEMO Years in USA</td>
<td>0.58</td>
</tr>
<tr>
<td>Conv. Frequency</td>
<td>DEMO Child’s Age</td>
<td>-0.54</td>
</tr>
<tr>
<td>Turn-Taking in Reading</td>
<td>PRQ Frustration</td>
<td>-0.52</td>
</tr>
<tr>
<td>Parent’s Speech Rate</td>
<td>DEMO English at Home</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

**TABLE 11: Significant correlations between sociodemographic and affect features ($|\rho| \geq 0.49$).**

<table>
<thead>
<tr>
<th>Affect Feature</th>
<th>Sociodemographic Feature</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent’s Arousal</td>
<td>DEMO Parent’s English</td>
<td>0.60</td>
</tr>
<tr>
<td>Parent’s Valence</td>
<td>DEMO Parent’s English</td>
<td>0.60</td>
</tr>
<tr>
<td>Child’s Engagement</td>
<td>DEMO Parent’s English</td>
<td>-0.51</td>
</tr>
<tr>
<td>Child’s Arousal</td>
<td>PSI Dysfunction</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

### 7.3 Pairwise correlations between sociodemographic and affect features

The results of the pairwise correlation analysis between the sociodemographic and affect features show that 4 of 102 correlations achieved the coefficient threshold ($|\rho| \geq 0.49$), as shown in Table 11. The results show that the parent’s affect is positively influenced by their use, exposure, and proficiency in English. Specifically, the correlation between their arousal and English Level is $\rho = 0.60$, while the correlation of their valence with their English Level is $\rho = 0.60$. This result indicates that parents who are more comfortable with English are more likely to exhibit higher arousal and more positive valence. Furthermore, the child’s engagement was found to be most correlated with the parent’s English level ($\rho = -0.51$). This result showed that children were less engaged if their parent’s English level was higher, probably because some children were engaged in reading stories more often if their parents were English learners instead of native English speakers. Lastly, the child’s arousal was correlated with PSI Dysfunction ($\rho = -0.50$). This result indicates that a parent’s parenting style and stress impact their child’s displayed affect during shared reading.

### 8 Research Direction Recommendations

This section presents the potential applications for the dataset in the fields of education, psychology, affective computing, and human-computer interaction. The limitation of the dataset is also discussed.
8.1 Investigating social-affective dynamics of parent-child interactions

The DAMI-P2C dataset provides many opportunities to investigate social-emotional dynamics of parent-child interactions in more depth using the different data modalities provided. First, future work can leverage the dataset to examine questions pertaining to multimodal social cues of parent-child dyads. A wide range of high-level dyadic verbal and nonverbal behaviors, such as gesture and head nodding, can be extracted and investigated, given the provided affect labels and body tracking features for every person. As the earliest attempt to utilize the DAMI-P2C dataset, we examined the body gesture and head movement indicators of the parent-child relationship and found their relationship characteristics associated with both individuals’ and dyads’ nonverbal behaviors holistically and relationally rather than in isolation. The work demonstrates the viability of using the DAMI-P2C to uncover parent-child dynamics and motivates future work in this field to use the dataset. Building upon our previous work in [91], one interesting future research direction is, for example, to examine how the parent’s and child’s social cues mutually adapt to each other during co-reading and how a participant’s social cues regulate the other’s affect in this context.

Additionally, the video data recorded by multi-view cameras allow dyadic multimodal social cues to be analyzed in the 3D space with depth perception analysis. For example, a wide set of 3D behaviors related to interpersonal space and body dynamics could be extracted, such as touching and fixing hair, patting the back, the number of bodies overlapping, and their depth distance. These 3D nonverbal behaviors could be further used to develop or evaluate behavioral assessment measures for dyadic interaction dynamics, given their links with psychological constructs, such as joint engagement [92], interpersonal synchrony [93], emotional availability [94], and bidirectional mutuality [95].

Furthermore, the dataset allows for in-depth analysis of individual similarities and differences of affective and reading behaviors when comparing across parent-child dyads. One interesting research avenue is, for example, to analyze how sociodemographic parameters of parent-child dyads can contribute to observed across-dyad variations in nonverbal, conversational, and affective behaviors by performing multivariate regression analyses. The results of such research could potentially assist psychologists to develop effective targeted interventions on parent-child social-emotional interactions. In summary, the dataset contributes to the research on uncovering the social and affective dynamics of parent-child interactions.

8.2 Developing recognition models of human-human dyadic interactions

The multimodal nature of the dataset enables the development of affect recognition models in a human-human dyadic interaction context. Potential affective modeling tasks supported by the dataset range from detecting adults’ and children’s valence, arousal, and engagement to predicting adult-child nonverbal communication styles and interaction dynamics. The dataset also supports multimodal, multi-sensor approaches to affective modeling, as it provides videos from multiple camera angles, audio, speech transcripts, and sociodemographic data. As the earliest attempt to utilize the DAMI-P2C dataset for affect modeling, Chen et al. [36] developed single-modality deep learning models trained on the speech data to recognize valence and arousal of individuals in dyadic conversations, and the results provide a competitive baseline for future research in this field. Future work can leverage the DAMI-P2C dataset in multiple novel directions.

First, the provided raw audiovisual recordings can be used to train end-to-end deep learning models for affect detection tasks. For example, the models in Chen et al. [36] took frame-level acoustic features as their input, but future work could take the raw waveform provided in the dataset as the model input directly by adding a time-convolutional layer [96], which can learn an acoustic model and transform the raw time-domain waveform to a frame-level feature. This end-to-end extension has the potential to reduce the affect-sensing system’s reliance on off-line pre-processing and enlarge its capability of sensing and reacting to human affect in real-time interactions.

Second, the availability of multiple affect labels annotated for both interlocutors supports the development of recognition models that can jointly learn individual’s affect labels. In addition to building multi-task affect models, leveraging multiple modalities in the dataset would further increase the applicability of affect detectors to challenging real-world cases, e.g., handling technical difficulties, such as background noise and poor lighting [39]. Furthermore, open-sourced audiovisual recordings would encourage future researchers to annotate and build affect models for a greater variety of affective states, e.g., mind wandering [97], [98]. Overall, the dataset could be used for developing and comparing a variety of state-of-the-art multimodal, multi-task affect recognition implementations, e.g., [99].

Third, DAMI-P2C allows for the exploration of building affect-sensing systems personalized to individual variations in affect across dyads. Modeling affect and interaction dynamics in a personalized and culture-sensitive manner has been shown to outperform one-size-fits-all affect models (e.g., [100], [101]). Individual differences also modulate affective expressions suggested by the theoretical foundation for affect detection [38], and cultural differences have been empirically found to exist in some aspects of emotions, particularly emotional arousal level [102]. Thus, personalized or culture-sensitive affect models may further improve model prediction performance. To build next-generation affect detectors, the social and cultural characteristics of individuals and dyads can be leveraged to build personalized culture-sensitive models. Participant families in DAMI-P2C came from diverse social and cultural backgrounds, and such individual differences are captured by the sociodemographic profiles in the dataset.

Overall, the dataset is a unique contribution to the research of modeling multiperson affect in human-human interactions. We acknowledge that the imbalance of affect annotation ratings in the dataset (as shown in Fig. 4) may add the difficulty of training automatic detection models if without applying any data resampling and/or augmentation or other measures (e.g., adjusting the loss function) to address the imbalance in data points. Given the prevalence
of imbalance of data labels in real-world data, plenty of techniques have been developed and widely used to address the issue of imbalanced datasets in machine learning for different feature types and modalities (e.g., image manipulation \[103\]. Generative Adversarial Networks for images and speech signal \[104\], resampling techniques for statistical and temporal data \[105\]). Furthermore, the performance of a model could be measured using metrics that account for this data imbalance, such as F1 and Matthews Correlation Coefficient (MCC).

8.3 Designing technology-based interventions on parent-child co-reading

Given its potential use in both investigating parent-child interaction dynamics and developing future multiperson affect recognition models, this dataset is also a valuable resource for informing the design of future interactive technologies, such as social robot companions in both single-person and multiplayer contexts.

First, the social and reading behaviors of parents in this dataset can be used to develop behavior policy models for interactive story-reading technologies for children, such as social robot companions. Dyadic human-human datasets have been widely used to guide the design and development of social agents that interact with single users. For example, Lee et al. \[106\] used the P2PSTORY dataset \[48\], a child-child dyadic storytelling dataset, where the human storyteller and listener have distinct roles, to train a Bayesian behavior policy for a social robot listener to communicate its attention in dyadic storytelling interactions with individual children \[106\].

Second, the behavioral analysis of parent-child dyads in DAMI-P2C can inform the design of interactive technologies aimed at promoting parent-child dyadic interactions. For example, it can provide insights into when parent-child disengagement and behavioral asynchrony may occur and how they are handled in co-reading. It was observed in the dataset that parent-child dyads often lost joint engagement or behavioral synchrony for a certain period of time when engaging in the co-reading activity. However, some parent-child dyads very quickly recovered by adapting how they communicate with each other, while others struggled to keep the dyadic experience uplifting for the rest of the activity. The handling strategies employed by fast-recovering dyads could be integrated into interactive technologies to assist the recovery of struggling dyads. Furthermore, the rich modalities of DAMI-P2C allow the designing of technologies’ intervention strategies to be personalized, such as intervening through a set of distinct verbal and nonverbal modalities, e.g., speech and text, based on each family’s preference.

Overall, the dataset makes a unique contribution to the design of both child-technology dyadic interactions and parent-child-technology triadic interactions in the early childhood education context. One of the fundamental bases of human-computer interaction research stems from Reeves’ and Nass’ work demonstrating how the human mind will respond to technology as social actors capable of evoking the same social responses as they would with a human partner \[107\]. Through data-driven methods to better understand the social-emotional dynamics of human dyadic interactions, we can more appropriately design human-like behaviors of learning technologies.

8.4 Dataset Limitations and Recommendations

We acknowledge that the dataset might be considered small compared with others in terms of the total number of recruited participants. However, the richness and variance of the features provided in DAMI-P2C can be used to conduct impactful quantitative analyses or build machine learning models. Prior works exist in demonstrating how to build interaction models using small sample sizes \[108\], \[109\]. For example, the P2PSTORY dataset, a child-child storytelling dataset with 18 participants and a total of 75 minutes of audiovisual interaction data, has been used to train an inference model for nonverbal communication modeling, e.g., \[106\], and demonstrated the importance of nonverbal features in predicting backchanneling extent and listener disengagement \[110\]. Similarly, prior work also demonstrated how a small participant size dataset could be useful for developing technology-based interventions (e.g., \[28\]). For example, in \[28\], a small dataset with fewer than ten participants, was used to train the initial parameters of an intelligent agent’s behavior model, helping accelerate the model’s learning and training in an intervention experiment. Hence, we believe that our dataset can also be a powerful resource for investigating parent-child interaction dynamics, developing recognition models of human-human interactions, and designing technologies that interact with children and parent-child dyads.

Besides the provided participant-level data, such as dyad’s sociodemographic profiles, the behavior-based data displayed or extracted within interaction sessions, e.g., affect labels and nonverbal cues of participants, indeed provide a large size sample size (e.g., the total number of annotated affect segments is 16,593). When used on the level of nonverbal or affective data instances, the dataset has been demonstrated in prior work to be useful and effective in yielding analysis insights on human-human interaction dynamics (e.g., studying nonverbal indicators of dyadic relationship characteristics \[21\]) and training state-of-the-art models (e.g., end-to-end deep learning models for speech-based multiperson affect recognition \[35\]). Future research could also analyze the interrelations between dyad’s affective states and nonverbal behaviors, such as affective triggers of certain nonverbal behaviors, and vice versa.

Lastly, we acknowledge that the preliminary analysis in this paper is constrained by the number of participants in the dataset (n = 34), which might limit the generalizability of specific results to broader parent-child interaction phenomena and their interrelations with the parent-child relationship styles. However, this analysis is mainly intended to describe the statistic characteristics of the dataset and show the interrelations among available data modalities. In other words, the analysis is exploratory rather than confirmatory. In the future, more in-depth statistical analyses, e.g., regression models, are needed to pinpoint specific interrelations among features from different data modalities. In summary, the presented preliminary analysis serves as the first step toward understanding and assessing the parent-child relationships and/or interaction dynamics as well as their nonverbal and affective behaviors.
9 CONCLUSION

DAMI-P2C is a multimodal parent-child interaction dataset built using 3-7 years old children. The objective of this dataset was to capture children and parents engaged in natural co-reading interactions, along with a diverse collection of their sociodemographic profiles. In addition, the dyadic reading behaviors, perceived affect of individuals, and 2D and 3D body joints of individuals were either extracted or annotated in the dataset, allowing for future in-depth investigation of parent-child interaction dynamics, the development of multiperson affect recognition systems, and the design of interactive learning technologies.

We hope that this work will increase awareness of the need for such studies collecting data on parent-child interactions as well as provide useful insights for both the computing field (e.g., affective computing and human-computer interaction) and the social science field (e.g., psychology and education).

10 ACKNOWLEDGEMENTS

We thank Xiahjie Zhang for helping conduct user studies. We also thank Michaela Markwart, Hollin Hakimian, Kaithlin Zareno, and Aditi Abhayakumar for annotating videos. This work was supported by the Intel Graduate Fellowship and the MIT Integrated Learning Initiative (MITili) grant.

REFERENCES

discussio...s. User Modeling and User-Adapted Interaction, 26:393–423, December 2016.


