Towards Automated Annotation of Infant-Caregiver Engagement Phases with Multimodal Foundation Models

Daksitha Withanage Don
Dominik Schiller
Tobias Hallmen
Silvan Mertes
Tobias Baur
Florian Lingenfelser
Elisabeth André
University of Augsburg
Germany
daksitha.withanage.don@uni-a.de
dominik.schiller@uni-a.de
tobias.hallmen@uni-a.de
silvan.mertes@uni-a.de
tobias.baur@uni-a.de
florian.lingenfelser@uni-a.de
elisabeth.andre@uni-a.de

Mitho Müller
Lea Kaubisch
Corinna Reck
Ludwig Maximilian University Munich
Germany
mitho.muller@psy.lmu.de
lea.kaubisch@psy.lmu.de
corinna.reck@psy.lmu.de

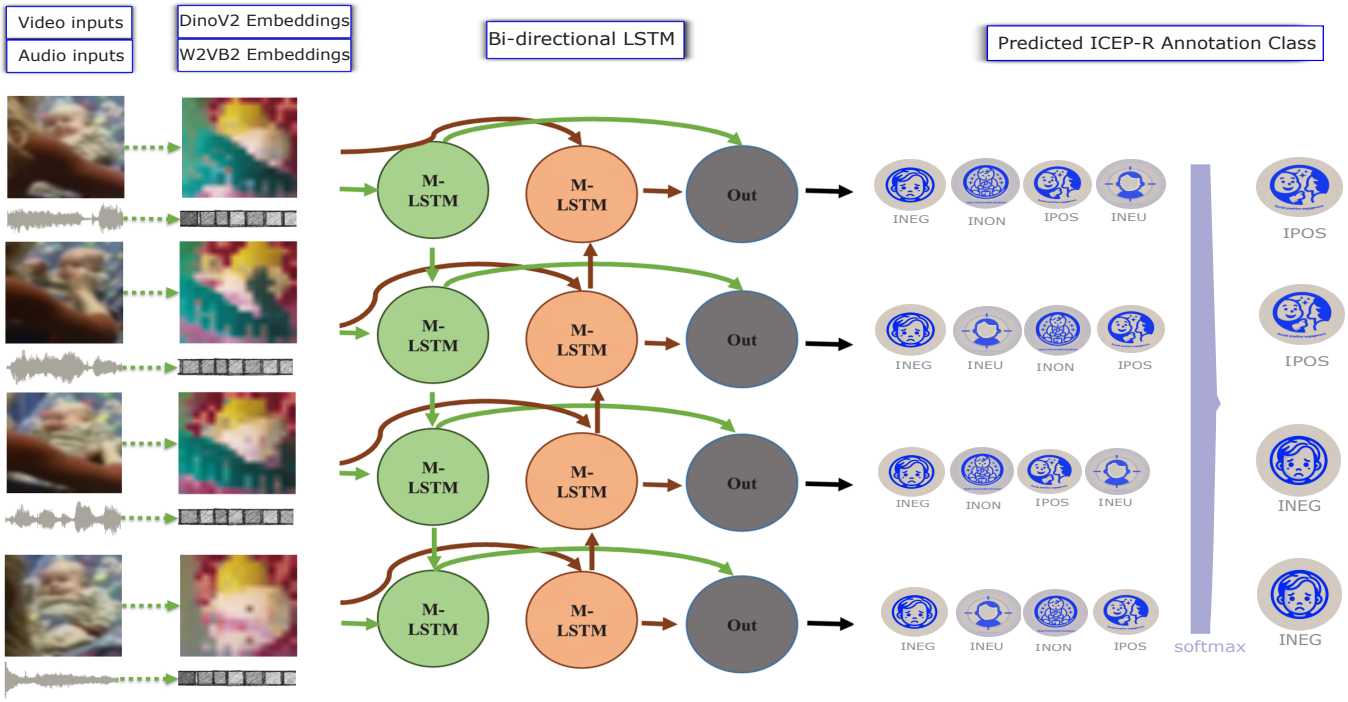


Figure 1: Inference calculation of the infant multimodal annotation predictor using bi-LSTM. DinoV2 PCA components (16 x16 x 4) are visualized in CMYK color space. Wav2Vec2-BERT is the size of 1024 audio feature embeddings.

Abstract

Caregiver mental health disorders increase the risk of insecure infant attachment and can negatively impact multiple aspects of child development, including cognitive, emotional, and social growth.

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685704

Infant-caregiver interactions contain subtle psychological and behavioral cues that reveal these adverse effects, underscoring the need for analytical methods to assess them effectively. The Face-to-Face-Still-Face (FFSF) paradigm is a key approach in psychological research for investigating these dynamics, and the Infant and Caregiver Engagement Phases revised German edition (ICEP-R) annotation scheme provides a structured framework for evaluating FFSF interactions. However, manual annotation is labor-intensive and limits scalability, thus hindering a deeper understanding of early developmental impairments. To address this, we developed a computational method that automates the annotation of caregiver-infant interactions using features extracted from audio-visual foundational models. Our approach was tested on 92 FFSF video sessions. Findings demonstrate that models based on bidirectional LSTM and linear classifiers show varying effectiveness depending on the role and feature modality. Specifically, bidirectional LSTM models generally perform better in predicting complex infant engagement phases across multimodal features, while linear models show competitive performance, particularly with unimodal feature encodings like Wav2Vec2-BERT. To support further research, we share our raw feature dataset annotated with ICEP-R labels, enabling broader refinement of computational methods in this area.

CCS Concepts

• **Computing methodologies** → **Neural networks**; *Behaviour analysis*; • **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → **Life and medical sciences**.

Keywords

Developmental psychology, Caregiver-infant interaction, Still Face Paradigm, Automated annotation, Self-supervised learning

ACM Reference Format:
Daksitha Withanage Don, Dominik Schiller, Tobias Hallmen, Silvan Mertes, Tobias Baur, Florian Lingenfelser, Elisabeth André, Mitho Müller, Lea Kaubisch, and Corinna Reck. 2024. Towards Automated Annotation of Infant-Caregiver Engagement Phases with Multimodal Foundation Models. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3678957.3685704

1 Introduction

From the earliest days of life, the intricate dance between infant and caregiver shapes the foundation for a lifetime. These interactions are complex and contain multiple aspects, which directly shape regulatory, emotional, cognitive, and social capacities in infants' development [12, 31]. Early theoretical frameworks, like Bowlby's trilogy on attachment [5, 6], and Ainsworth's deep dive into maternal sensitivity [1], emphasize the important role of responsive caregiving in developing secure bonds between caregiver and their infants. Later, research found that these caregiver-infant interactions are not one-directional. Rather, infants actively engage in these exchanges, influencing their developmental trajectories [38, 39]. Reck et al. [35] evaluated the impact of maternal anxiety disorder on caregiver-infant interaction in the postpartum period. Therefore, it is crucial for psychologists to understand and evaluate these behaviors for early interventions.

Developmental psychologists employ structured observation techniques to evaluate infant-caregiver interactions, including free play, standardized play, and the Face-to-Face Still-Face (FFSF) paradigm [26]. Our research focuses on the FFSF method, a standardized procedure designed to analyze infants' reactions to socio-emotionally challenging situations. This paradigm allows the evaluation of infant interaction patterns with behaviors exhibited after a socio-emotional stressor. It consists of three episodes, each lasting two minutes. Initially, during the play episode, the caregiver engages in typical play with the infant without the use of toys or pacifiers. This is followed by a transition interval where the caregiver turns their head aside and quietly counts to ten. The still-face episode then ensues, during which the caregiver maintains a neutral expression, refraining from any gestures or vocalizations, thereby creating a state of interactive mismatch. Finally, the reunion episode begins, with the caregiver resuming face-to-face play with the infant without a transition interval. Throughout these episodes, the interactions are video recorded and coded [41, 42].

In coding the interactions observed during the FFSF paradigm, two well-known coding techniques are employed: macro-temporal and micro-temporal behavior annotations. The Coding Interactive Behaviour (CIB) system [13] is a prominent example of macro-temporal annotation, where entire sessions are recorded on audio and video, and evaluated as complete sessions by CIB annotators across a range of behavioural characteristics. In contrast, micro-temporal annotation involves a more detailed approach, breaking down the behavioral stream into distinct intervals, often second by second. The Infant Carer Engagement Phases (ICEP)[46] and updated German version (ICEP-R) [34] illustrate this method, coding each stage of infant-caregiver involvement by assigning codes to each frame.

Although observational tasks and evaluation methods in infant caregivers have substantially advanced in the last few decades, they are not without their challenges. Translating qualitative observations into quantitative coding introduces an element of subjectivity, which can affect inter-rater reliability and the validity of findings across different research settings. To mitigate this, researchers often rely on rigorous training programs to standardize the coding process, but these programs are both time-consuming and costly. Aside from that, coding video sessions remains labor-intensive and resource-intensive, posing challenges for clinicians in developing regions that have limited access to such facilities. As a result, achieving equitable evaluation in these areas remains an ongoing challenge, underscoring the need for more efficient and accessible approaches to analyzing behavioral data.

The rapid progress of machine learning and neural networks holds great promise for analyzing intricate behavioral patterns in caregiver-infant interactions. Despite this potential, few studies have applied machine learning for micro-annotation analysis. [22] developed a model predicting attachment types (*secure* or *insecure*) using data from 64 infant-caregiver pairs during the FFSF and Ainsworth's Strange Situation Assessment (SSA). Their modest prediction accuracies for a two-class classification problem underscore the need for advanced computational methods to interpret these complex interactions.

In this work, we introduce a novel approach to automating the annotation process of ICEP-R for both infants and caregivers utilizing linear classifiers model and bidirectional LSTM model, as illustrated in Fig. 1. We wish to streamline the traditionally labor-intensive annotation task by employing standardized feature extraction techniques and deploying trained models to predict ICEP-R annotations on new data. By leveraging audio and video foundational backbone models to extract features instead of traditional ones like Mel-Frequency Cepstral Coefficients (MFCC) for audio and Convolutional Neural Networks (CNN) features for visual data, our method better captures holistic feature representation from audiovisual signals. Unlike previous studies that focused solely on binary attachment classification, we analyze seven distinct engagement types for infants and caregivers. To facilitate progress in this interdisciplinary field, we publicly release our pre-trained models, raw feature datasets, and software pipelines for researchers to access at https://github.com/Daksitha/SCHWAN-ICEP-R-Automation.git.

2 Related Work

Caregiver-infant interaction has been extensively researched in developmental psychology due to its critical role in early childhood development [3, 40]. Studies highlight the complex nature of early interactions between caregivers and infants, emphasizing their significance in creating a nurturing social learning environment [4, 14, 44]. These interactions are fundamental not only for cognitive and social learning but also for developing emotional regulation and establishing a secure caregiver-infant attachment [15, 16, 29]. Maternal mental health issues like postpartum depression and anxiety can have prevalence rates of up to 27.8% (EPDS $\geq$9) and 9.0% (EPDS $\geq$14), according to recent studies. SCID-based assessments also report a 9.0% prevalence, underscoring the critical importance of analyzing their impact on infant-caregiver interactions [25, 35].

The ICEP micro-annotation scheme, initially developed by Weinberg and Tronick [46] to analyze interactions between infants and their caregivers, was later refined into the ICEP-R system by [34] for German studies. Macro-temporal schemes such as CIB [13], the Attachment Q-Set (AQS) [45], and the Parent-Child Early Relational Assessment (PCERA) [10] are utilized to evaluate broader interaction patterns and the overall quality of relationships over extended periods. In our work, we have chosen to predict the ICEP-R annotation due to its micro-temporal granularity that matches the frame-level detail provided by audiovisual signals.

2.1 Current Computational Approaches in Caregiver-Infant Analysis

Recent advancements in computational methods have significantly enhanced the analysis of social interactions in developmental psychology, particularly in understanding infant-caregiver dynamics. This section reviews notable studies that leverage these methods to gain insights into parent-child interactions. Leclere et al. [21] utilized 2D and 3D Microsoft Kinect motion capture data to study mother-infant interactions. They correlated motion-derived parameters, such as shoulder angles and body distance, with macro CIB scores. A Support Vector Machine (SVM) was employed to effectively distinguish between high-risk and low-risk groups. However, the reliance on Kinect landmark-based features and CIB annotation schemes may limit the analysis by potentially overlooking subtle behaviors. Klein et al. [20] examined the coordination across various modalities—such as head and arm movements combined with vocal frequencies—to enhance the understanding of mother-infant interactions during the FFSF procedure. While this study successfully highlights the importance of multimodal analysis, its dependence on key-point and landmark-based features is prone to error in dynamic and complex environments like infant-caregiver interactions. Mills, Koonce, and Cox applied machine learning techniques to automate the evaluation of "three-bag-assessment" based parent-child interaction video recordings, aiming to enhance attachment-based interventions. They utilized OpenPose for pose estimation and vocal fundamental frequency of mothers and their infants for behavior assessment, correlating these with manual CIB ratings. Despite demonstrating potential for AI in streamlining interventions, the study's accuracy was limited by environmental variables such as camera angles and acoustics, highlighting the challenges of applying these technologies in varied settings. However, self-supervised model-based features offer versatility in such dynamic and challenging conditions, potentially mitigating the impact of environmental factors on automated assessment accuracy [27]. Despite the advancements, all three studies face limitations due to their reliance on manually crafted features and landmark-based annotations.

Li et al. [22] developed a model to predict attachment types as secure or insecure using 64 infant-mother pairs during the FFSF and Ainsworth's SSA. Instead of using landmarks for video analysis, they utilized video frames together with VGG9 to extract visual features, and audio features such as pitch, short-time energy (STE), and MFCC. They employed SVMs to classify the labels from audio and then used late fusion with VGG9 classification to predict the final label. Their approach showcased the potential of neural networks in identifying attachment patterns, achieving classification accuracies of 38% for the Still-Face phase, 60% for the Reunion phase, and 61% when combining both phases using only infant data. Incorporating caregiver data somewhat improved results to 50% (Still-Face phase), 67% (Reunion phase), and 78% (combined phases). Moreover, the authors mentioned the need to consider selecting context-based representative features and making a better fusion of visual and audio features. They also suggested expanding the dataset and further distinguishing among the insecure attachment types, including ambivalent attachment, avoidant attachment, and disorganized attachment. To address these issues, our work proposes the use of multi-labeled micro-annotation schemes and occlusion-invariant, holistic self-supervised learning-based feature representations for both audio and video signals. These methods aim to provide more detailed insights into the complexities of parent-child interactions, enabling computational models to classify between different attachment types more accurately.

2.2 Foundational Models and Temporal Modelling

Foundational models, trained on vast amounts of data, excel in learning a wide range of patterns and features, enabling them to generalize effectively across tasks and domains [23]. These models have achieved significant advancements in both Natural Language Processing (NLP) [9, 11, 18, 32] and computer vision [8, 30].

In NLP, speech encoder models like W2V-BERT 2.0 also known as Wav2Vec2-BERT [11] in SeamlessM4T v2 was pre-trained on 4.5 million hours of unlabeled audio and fine-tuned with supervised data, enhancing performance on low-resource languages. SeamlessExpressive, built from diverse datasets, preserves vocal style and prosody in translation. In computer vision, the DinoV2 [30] model demonstrates superior image classification performance by focusing on relevant information within images. For instance, DinoV2 achieves high accuracy on recognition benchmarks like UCF101 [37] (91.2%) but underperforms on temporally sensitive action prediction tasks such as SSv2 [17], with a 38.3% overall prediction accuracy using a linear classifier on averaged features from eight spaced frames. This discrepancy highlights the challenges in temporally sensitive action prediction. These challenges are amplified in contexts such as child-caregiver interactions, as the continuous and dynamic nature of behavioral annotations within a single video and audio session significantly increases the complexity of the modeling process.

Bidirectional Long Short-Term Memory (bi-LSTM) networks have achieved notable success in the domain of computer vision, particularly in the recognition of actions within video sequences [43]. Traditionally, actions in video sequences are considered static and can be described in detail, similar to a sentence that narrates the sequence of events in the video [17, 37]. However, the introduction of micro behavior annotations, such as those provided by ICEP-R, complicates this paradigm. Unlike conventional action labels, which remain constant for a given video, behavioral actions annotated with ICEP-R exhibit frequent changes within the same video, thereby increasing the predictive complexity compared to traditional action recognition tasks.

Addressing these challenges, our method evaluates the effectiveness of integrating bi-LSTM networks and linear classifiers with foundational model-extracted features for predicting ICEP-R annotations independently for both caregivers and infants. This process leverages the strengths of audio and video-based foundational models combined with both sequence modeling and traditional classification techniques to capture the intricate spatial and temporal dynamics of caregiver-infant interactions. To the best of our knowledge, this is the first instance of evaluating the combined use of bi-LSTM networks and linear classifiers with frozen audio-visual foundational model features to predict caregiver-infant engagement annotations.

3 Dataset

We obtained access to a video and audio dataset originally utilized to analyze the impact of postpartum anxiety disorder on caregiver-infant interactions and infant development [28, 35]. The dataset comprises video and audio recordings during FFSF interactions. FFSF episodes are *play* (i.e., unstructured interaction between infant and caregiver) *still face* (i.e. caregiver suddenly becomes unresponsive, observing the infant's reaction,) and *reunion* (i.e. normal interaction resumes after the still face episode, gauging the infant's recovery).

The dataset includes a total of 92 video and audio sessions. The participants were divided into two groups: 39 women diagnosed with DSM-IV postpartum anxiety disorder and 53 healthy control caregivers. However, the diagnostic information is not considered in our predictive task. The infants involved had an average age of 4.1 ± 1.5 months at the time of the study. The caregivers in the study had an average age of approximately 33 years and all are female. 58% of the caregivers held a university degree, indicating a high level of education among the participants. All participants were Caucasian ranging from German and French nationalities, and 51% were in a marital or stable relationship at the time of the study. caregiver and the infant interact with each other for about 6 min [28], as illustrated below.

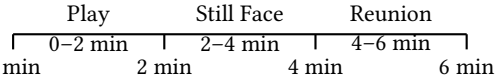


Table 1: Overview of Infant Engagement Phases

Phases	Modality	Explanation
Negative Engagement (**Ineg**)		Further divided into **Ipro** + **Iwit**
Protest (**Ipro**)	Visual + Audio	Active negative behaviors, e.g., crying, fussy vocalizations, arching back or kicking.
Withdrawn (**Iwit**)	Visual	Shows sadness, lack of focus, gaze aversion, minimal movement.
Object/Environment Engagement (**Inon**)	Visual	Focus on objects, interest or neutral expressions, may vocalize.
Social Monitor (**Ineu**)	Visual + Audio	Neutral/interested gaze at caregiver, neutral/positive vocalizations.
Social Positive Engagement (**Ipos**)	Visual + Audio	Positive interactions, joy, engaging in play, specific facial cues.
Sleep (**Islp**)	Visual	Infant is asleep, no interaction.
Unscorable (**Iusc**)	Visual	Cannot be scored due to obscured view or partial visibility.

3.1 ICEP-R Annotations

Initially, the ICEP coding system, by Tronick et al. [46], was translated and revised into German by Reck et al. [34]. These 6-minute video sessions are then analyzed and coded using the ICEP-R coding system [28]. This approach involves detailed coding of each time interval to categorize behaviors into distinct phases for both the infant and the caregiver. The videos are coded by two trained and certified, blinded coders using split video, where infant and caregiver videos are synchronously merged side by side to create a single video with a unified audio track. Annotators used the Noldus Observer Video-Pro system, with interrater reliability for engagement phase codes measured by Cohen's κ. ($\kappa = [0.73; 0.82]$ for infant codes; $\kappa = [0.72; 0.73]$ for maternal codes). It is important to note that the codes within the infant and caregiver categories are mutually exclusive, ensuring that there is no overlap in the coding of behaviors.

Tables 1 and 2 provide a summary of the detailed annotation scheme used by expert annotators. Table 1 outlines the behavioral categories for infants, detailing the specific actions and expressions considered during the coding. Table 2 presents the corresponding behavioral categories for caregivers, focusing on their interactions and responses to the infants and their surroundings. Both tables highlight the mutually exclusive nature of the codes and the specificity required in the annotation process. Additionally, Fig. 2 visually

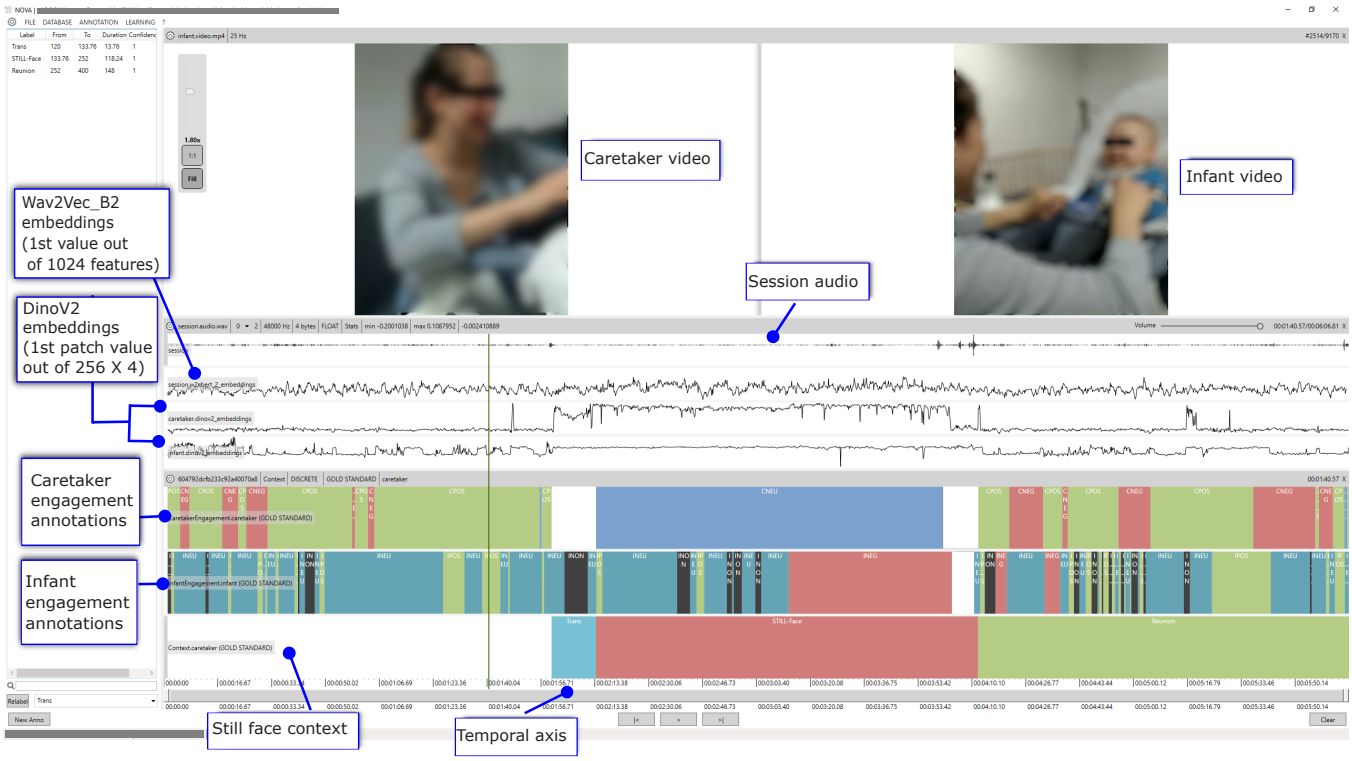


Figure 2: Overview of Infant-Caregiver FFSF on NOVA [2, 19]

Table 2: Overview of Caregiver Engagement Phases

Phases	Modality	Explanation
Negative Engagement (**Cneg**)		Further divided into **Cwit** + **Cint** + **Chos**
Withdrawn (**Cwit**)	Visual + Audio	Minimally engaged, withdrawn; sad, flat, expressionless face; silent, speaks in monotone.
Intrusive (**Cint**)	Visual + Audio	Intrusive behavior; tense, stressed expressions; does not wait for infant reactions, interrupts infant.
Hostile (**Chos**)	Visual + Audio	Hostile affect; stressed, aggressive expressions; high-pitched vocalizations; curt with baby.
Non-Infant Focused Engagement (**Cnon**)	Visual + Audio	Not attending to baby; involved in other activities like fixing clothing or talking to others.
Social Monitor/No Vocs or Neutral Vocs (**Cneu**)	Visual + Audio	Watches baby, neutral expressions; may touch baby; vocalizations are neutral if present.
Social Monitor/Positive Vocs (**Cpvc**)	Visual + Audio	Focused on infant; neutral, interested expressions; positive vocalizations, including motherese.
Social Positive Engagement (**Cpos**)	Visual + Audio	Expresses positive affect; smiles, laughter; playful interactions; uses motherese or sings.
Unscorable (**Cusc**)	Visual + Audio	Cannot score due to obscured view; unclear vocalizations; face partially or completely hidden.

illustrates these annotations completed for a session, aiding in the understanding of the coding scheme.

4 Method

Figure 3: Left: Detection of face and body keypoints under occlusion (marked by distinct nodes and connections). Right: Visualization of the first four principal components (different regions) of DinoV2 attention map, represented in a CMYK color space.

Our primary objective is to develop predictive models to predict infant and caregiver ICEP-R annotations, as illustrated in Fig. 4, from multimodal audio-video data. The problem can be formulated

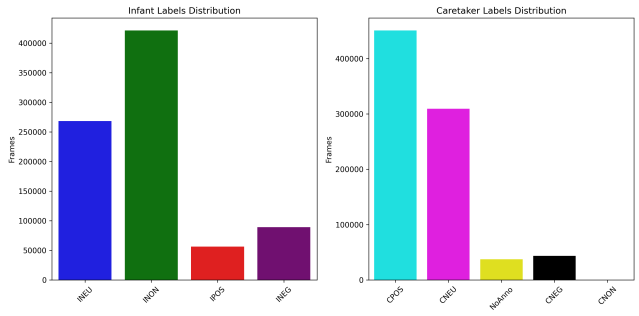


Figure 4: An overview of infant-caregiver label distribution for all sessions

as follows: given a sequence of features from audio and/or video data, the model should predict the behavioral annotation class to which these feature frames belong.

Initially, we extracted facial and body landmarks with Mediapipe holistic [24] and face-alignment [7], illustrated in Fig. 3 (left), which can be misidentified, especially when occlusions occur in video images. Additionally, audio data often includes background noise, complicating the identification of key behavioral indicators and impeding accurate classification of ICEP-R annotations. To address these issues, we propose a feature extraction process using self-supervised foundational models for both audio and video. These feature embeddings provide richer representations and are invariant to background noise. For example, in the same video shown in Fig. 3, caregiver occlusions are mostly present during the *play* and *reunions* phases and less so during the *still-face* phase. Vision self-supervised learning methods, particularly those using attention mechanisms, can focus on regions of interest without altering model parameters during these behavioral changes, as illustrated in Fig. 3 (right).

4.1 Feature Extraction

We have integrated our feature extraction pipeline into NOVA, an open-source toolkit for annotating and analyzing behaviors in social interactions [2, 36]. NOVA enables visualization of multiple synchronized media tracks, including video and audio, along with discrete and continuous annotation tracks, as shown in Fig. 2. It allows for the integration of custom Python-based addons. We developed modules for foundational model-based feature extraction using DinoV2 for the video and Wav2Vec2-BERT (W2V2B) embeddings for the audio. We plan to integrate trained automated annotation of ICEP-R annotation into NOVA, enabling non-experts to use it with a user-friendly graphical user interface.

Our DinoV2 feature extractor module in NOVA:. The vision foundational model DinoV2 processes an image by producing a class token and patch tokens, with an optional inclusion of four register tokens. The embedding dimensions for this model are 384 for ViT-S, 768 for ViT-B, 1024 for ViT-L, and 1536 for ViT-g. DinoV2 employs a Transformer architecture with a patch size of 14. For a 224×224 image, this configuration yields one class token, 256 patch tokens, and optionally, four register tokens. The model can accommodate larger images, provided the image dimensions are multiples of the patch size (14); otherwise, it crops the image to the closest smaller multiple of the patch size [30].

We utilized the large model (ViT-L) due to its better quality in generating attention maps for our video data, resulting in a feature embedding size of 1024 dimensions per patch. As per Eq. 1, ViT-L outputs SE_{dinoV2} features for a video session with N frames. These features, extracted from video frames, are exponentially large and not directly interpretable when visualized as 16 by 16 by 1024 inferno-map images. Therefore, we applied Principal Component Analysis (PCA) to select the most prominent components, aligning the feature dimensions between audio and video embeddings. The number of PCA components selected, as indicated by Eq. 2, ensures a balanced early fusion of embeddings from both modalities. Visualization of these PCA feature embeddings, shown in Fig. 3, depicts the first four PCA components plotted in CMYK color space. Additionally, the first three PCA components can be plotted in RGB color space for further analysis.

$$SE_{dinoV2} = (N \times 256) \times 1024 \tag{1}$$

After applying PCA, the reduced feature dimensions are:

$$SE_{dinoV2_{PCA}} = (N \times 256) \times n\_compnents \tag{2}$$

We selected the first four PCA components to align the dimensions with the audio features and then reshaped the data to provide a per-video-frame feature vector. The dimensions of these feature embeddings are given by Eq. 3:

$$SE_{dinoV2_{PCA_4}} = N \times (256 \times 4) = N \times (16 \times 16 \times 4) = N \times 1024 \tag{3}$$

Given 92 video sessions, the size of the DinoV2 extracted feature embeddings is represented as $E_{dinoV2_{PCA_4}} = 92 \times SE_{dinoV2_{PCA_4}}$. The dimensionally reduced features are stored in a NOVA-compatible continuous annotation stream, enabling users to visualize these features alongside annotations and video and audio streams, as illustrated in Fig. 2.

Furthermore, we observed emerging patterns in the DinoV2 features that align with the *play*, *still-face*, and *reunion* (FFSF) phases. Figure 5 illustrates the computed mean values using the first four PCA features, revealing distinct patterns across multiple sessions. These patterns highlight transitions between the free-play (0-2 min), still-face (2-4 min), and reunion (4-6 min) phases. Notably, these patterns are inversely reflected in the feature streams for both the caregiver and infant. The implications of these observations are further discussed in Section 6.

Our W2V2B audio feature extractor module in NOVA:. The audio foundational model W2V2B process any length of audio by transforming it into a series of frame-level embeddings. This model operates with a fixed sampling rate of 16 kHz, ensuring consistent input across various audio sources. The embedding dimensions for W2V2B are configurable based on the specific model variant used, with typical values being 768 for base models and 1024 for large models [11].

In our audio feature extractor, a sliding window mechanism with adjustable stride left context (C_l), and right context (C_r) divides the continuous audio signal into frames. This flexibility allows for precise synchronization with video frames by incorporating relevant past and future contexts. For a given session audio signal

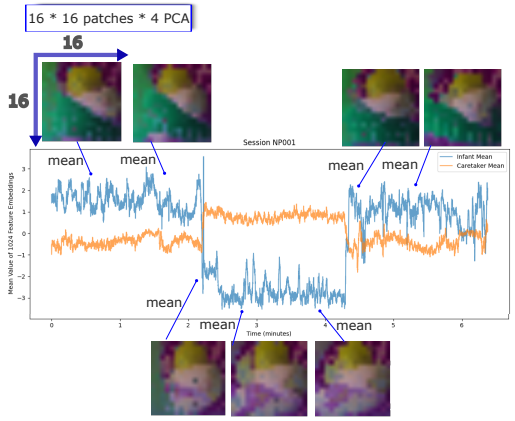


Figure 5: Emerging correlations between DinoV2 vs FFSF.

$SE_{W2V2B}(t)$, the window at the n^{th} video frame $SE_{W2V2B_w}(t_n)$ captures acoustic features essential for unified audio-visual alignment, which is crucial for our predictive task.

$$SE_{W2V2B_w}(t_n) = SE_{W2V2B}(t + n \cdot \Delta t - C_l, t + n \cdot \Delta t + C_r) \quad (4)$$

For synchronization with video frames, in Eq 4, we set $C_l = 50$ ms and $C_r = 50$ ms, while the video is sampled at 25 fps, resulting in $\Delta t = 40$ ms. These embeddings are then averaged over time to produce a 1024-dimensional vector per frame, representing corresponding audio features. The number of video samples in the session N matches between audio $SE_{W2V2B} = N \times 1024$ and video (Eq 3), ensuring synchronized multimodal analysis. Similar to the DinoV2 module, these saved embeddings can be visualized in NOVA for further inspection, as shown in Fig. 2.

4.2 Model Training

Once the feature embedding datasets for video and audio were created, we prepared the annotations as prediction labels by assigning each feature embedding group with the corresponding coding. As shown in Fig. 2, these annotations are discrete labels that span specific time segments. To address the imbalance and scarcity of certain annotations, we combined similar categories: Ipro and Iwit were grouped as Ineg, and Cwit, Cint, and Chos were grouped as Cneg. The distribution of these annotation labels is illustrated in Fig. 4. The NoAnno category represents segments that do not belong to any annotatable category. For classification tasks, we excluded the NoAnno category and defined a 4-class classification problem for infants and a 3-class classification problem for caregivers, excluding Cnon due to its minimal representation with only 36 labels. To prevent data leakage, we ensured that the data splits for training and testing were based on session names. This approach guaranteed that feature embeddings from the same video and audio sessions were not mixed between the training and testing sets. To evaluate the performance of predictive tasks using the aligned feature embeddings, we employed a dual approach. Initially, we explored the capabilities of a linear classification model. Subsequently, we extended our analysis by incorporating a more advanced Bi-LSTM model.

Initially, we wanted to determine the effectiveness of a linear classifier in capturing the intricacies of this predictive task. We developed and trained a linear classifier and evaluated using a single-run train-test split approach, allowing us to establish a baseline performance for the linear method across different modalities and for both infants and caregivers. For a more nuanced understanding of the temporal dependencies inherent in sequential data, we implemented a bi-LSTM classifier. This architecture was specifically chosen to leverage the LSTM's ability to model both forward and backward temporal dependencies, thereby capturing the dynamics of behavioral cues present in caregiver-infant interactions. Further details about linear classifier and bi-LSTM model hyperparameter tuning and the algorithm can be found in Appendix A.

5 Results

Table 3: Caregiver (3 class): Performance metrics using bi-LSTM (K-fold cross-validation) and linear classification (single run) across different modalities. W indicates weighted and UW for unweighted.

Feature Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (UW) (%)	F1 (W) (%)
Multimodal	bi-LSTM	72	87	72	62	76
	Linear	71	74	71	53	72
DinoV2	bi-LSTM	80	87	80	68	**82**
	Linear	68	68	68	49	66
W2V2B	bi-LSTM	63	84	63	51	69
	Linear	81	81	81	56	81

Table 4: Infant (4 class): Performance metrics using bi-LSTM (K-fold cross-validation) and linear classification (single run) across different modalities. W indicates weighted and UW for unweighted.

Feature Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (UW) (%)	F1 (W) (%)
Multimodal	bi-LSTM	68	72	68	60	**68**
	Linear	54	50	54	40	51
DinoV2	bi-LSTM	66	72	66	57	66
	Linear	47	41	47	26	43
W2V2B	bi-LSTM	55	56	55	43	50
	Linear	56	50	56	41	51

We analyzed the performance of bi-LSTM (using K-fold cross-validation) and linear classification (single run) models on the ICEP-R annotation task for caregiver and infant data across various feature modalities: multimodal, DinoV2, and W2V2B.

The results for the caregiver dataset, as presented in Table 3, indicate that the bi-LSTM model marginally outperformed the linear model when utilizing multimodal features, with a weighted F1 score of 76% compared to 72% for the linear model. Notably, when employing DinoV2 features, the bi-LSTM model demonstrated a significant performance advantage, achieving a weighted F1 score of 82%, underscoring its efficacy in capturing temporal information. In contrast, the linear model excelled with W2V2B features, attaining a weighted F1 score of 81%, thereby outperforming the bi-LSTM model.

Similarly, for the infant dataset, as shown in Table 4, the bi-LSTM model outperformed the linear model when using both multimodal and DinoV2 features, although the performance gap was narrower compared to the caregiver data. However, with W2V2B features,

the linear model's performance was comparable to that of the bi-LSTM model, suggesting that the linear model may be more suited to W2V2B features, which inherently contain temporal information within the embeddings. This contrasts with DinoV2 features, which, being extracted at the image level, do not inherently capture temporal information, thus favoring the bi-LSTM model. Further information of the K-fold cross-validation results can be found in Appendix B.

6 Discussion

These findings highlight several key insights. First, the bi-LSTM model generally demonstrates overall superiority in handling complex features, particularly in multimodal and DinoV2 datasets across both roles. In contrast, the linear classification model shows competitive or even superior performance when using W2V2B features, suggesting that simpler, linear methods may be more effective in certain feature spaces. Additionally, the performance of both models varied significantly between the caregiver and infant roles, with the bi-LSTM model showing more consistent performance across different roles and features.

This variability in model performance across different modalities can be partially attributed to our primary data source, the Face-to-Face Still-Face (FFSF) sessions, which provided a unique audio-visual perspective on caregiver-infant interactions. During these sessions, caregivers dominated the verbal exchanges, while infants, aged 3-5 years, primarily produced non-verbal sounds like crying, giggling, and occasional utterances. As the audio was captured on a single channel, both infant sounds and caregiver speech contributed to the W2V2B features. The linguistic content was atypical, consisting of repetitive phrases and playful sounds, often in multiple languages such as German and French. These factors likely influenced the distinct performance of the models across different feature types and roles.

Improving prediction accuracies may require addressing the challenges posed by the distinct characteristics of the audio data in this study. Models that rely exclusively on W2V2B features, which are primarily optimized for spoken word recognition, may not fully capture the diverse audio content typical of caregiver-infant interactions. This limitation likely contributed to the observed lower performance when using audio embeddings between infant and caregiver, which in turn affected the overall performance of the multimodal classification. To mitigate this issue, fine-tuning W2V2B with FFSF-specific data could enhance the model's ability to process the unique audio features present in these sessions, potentially leading to higher accurate predictions.

In addition to improving prediction accuracy through fine-tuning W2V2B, the integration of DinoV2 features offers distinct advantages in infant ICEP-R classification. Visual inspection of several DinoV2 PCA components revealed their effectiveness in capturing detailed movement patterns, including the infant's upper and lower motor activities and the caregiver's presence or absence during the still-face phase, as illustrated in 5. Unlike previous studies that relied on manual measurements or costly setups like Microsoft Kinect [21] to assess distances and correlate them with behavioral annotations, DinoV2 inherently encoded these features while also preserving the broader context of interactions. This led to an undemanding and accessible representation for analyzing and predicting caregiver-infant behaviors.

Furthermore, we explored emerging patterns from the FFSF sessions (Fig 5) by averaging the mean values across four DinoV2 PCA components. The resulting graph showed higher mean values during the mother's presence, with a notable peak during the reunion phase, representing the mother-infant interaction. Conversely, the still-face phase exhibited a dip when the mother left the scene, capturing mainly the infant's movement and surrounding context. This phenomenon closely resembles Motion Energy Analysis (MEA) [33], a tool used for synchrony evaluation that evaluates pixel-level movement to understand interaction dynamics and synchrony. However, DinoV2 exceeds MEA by offering more contextual data, like capturing the interaction scene's surroundings and handling occlusions more skillfully. The results in the caregiver ICEP-R prediction tasks further demonstrated how complete DinoV2's feature representation was, which further enhanced our predictive modeling.

The multimodal strategy showed higher performance in the infant categorization task, likely due to the dynamic nature of infant behavior. Infants frequently shift between engagement phases, making visual signals alone insufficient for accurate classification. The inclusion of the caregiver's verbal input provided important context, enabling the model to detect patterns that might otherwise be missed. Additionally, the frequent occlusion of infants by caregivers made auditory cues critical for identifying behavioral patterns, complementing the visual data, and enhancing classification accuracy.

However, there are notable limitations to this approach. The dataset's uneven distribution of annotations could introduce biases, particularly in underrepresented behaviors. Moreover, the distinctiveness and linguistic variation of the audio data, along with lower audio quality and the lack of separate channels for infants and caregivers, may constrain the model's generalizability across different behavioral classes. Addressing these challenges will require further research and refinement of the data.

7 Conclusion and Future Work

In this study, we introduced a novel multimodal approach that integrates auditory and visual cues to automate the annotation of ICEP-R, focusing on capturing the diverse and subtle behaviors between infants and caregivers. By leveraging bi-LSTM temporal modeling alongside advanced self-supervised foundational models, we moved beyond simple binary classifications to provide a more nuanced analysis of these interactions. The use of DinoV2 features offered detailed insights, and our multimodal strategy improved classification accuracy, particularly in handling the dynamic nature of infant behaviors. Despite challenges such as audio quality, linguistic diversity, and data imbalance, our findings demonstrate that this approach is effective in predicting ICEP-R annotations.

For future work, we plan to fine-tune audio foundational models specifically for infant-caregiver sessions and work towards separating audio activities for each role. Given our observation that feature embeddings with temporal information tend to perform better with simpler linear classifiers, we aim to incorporate state-of-the-art

video foundational models, such as V-JEPA, instead of image-based models, to extract feature embeddings from the sessions.

Acknowledgments

This paper has been partially funded by the German Research Foundation (DFG) within the SCHWAN Project (project number: 490909448).

References

[1] Mary D. S. Ainsworth, Mary C. Blehar, Everett Waters, and Sally Wall. 1978. *Patterns of Attachment: A Psychological Study of the Strange Situation.* Lawrence Erlbaum, Oxford.

[2] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F. Valstar, Björn Schuller, and Elisabeth André. 2020. eXplainable Cooperative Machine Learning with NOVA. *KI - Künstliche Intelligenz* (19 Jan 2020). https://doi.org/10.1007/s13218-020-00632-3

[3] D. Benoit. 2004. Infant-parent attachment: Definition, types, antecedents, measurement and outcome. *Paediatric Child Health* 9, 8 (Oct 2004), 541–545. https://doi.org/10.1093/pch/9.8.541

[4] Marc H Bornstein and Catherine S Tamis-LeMonda. 2001. Mother-infant interaction. In *Blackwell handbook of infant development*, J. Gavin Bremner and Alan Fogel (Eds.). Blackwell Publishing, Malden, 269–295.

[5] John Bowlby. 1969. *Attachment and Loss, Volume 1: Attachment.* Hogarth, London. Reprinted 1982.

[6] John Bowlby. 1980. *Attachment and Loss, Volume 3: Loss.* Hogarth, London.

[7] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294 [cs.CV]

[9] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. *CoRR* abs/2108.06209 (2021). arXiv:2108.06209 https://arxiv.org/abs/2108.06209

[10] Roseanne Clark. 1985. Parent-Child Early Relational Assessment (PCERA).

[11] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *ArXiv*.

[12] Elisabeth Conradt and Jennifer Ablow. 2010. Infant physiological response to the still-face paradigm: contributions of maternal sensitivity and infants' early regulatory behavior. *Infant Behavior and Development* 33 (2010), 251–265. https://doi.org/10.1016/j.infbeh.2010.01.001

[13] R. Feldman. 1998. *Coding interactive behavior manual.* https://ruthfeldmanlab.com/coding-schemes-interventions/ Unpublished Manual; Bar-Ilan University, Israel.

[14] Ruth Feldman. 2007. Parent-infant synchrony and the construction of shared timing; Physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry* 48, 3-4 (2007), 329–354.

[15] Ruth Feldman. 2012. Parent–infant synchrony: A biobehavioral model of mutual influences in the formation of affiliative bonds. *Monographs of the Society for Research in Child Development* 77, 2 (2012), 42–51.

[16] Ruth Feldman, Charles W Greenbaum, and Nurit Yirmiya. 1999. Mother–infant affect synchrony as an antecedent to the emergence of self-control. *Developmental Psychology* 35 (1999), 223–231.

[17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. *Proceedings of the IEEE International Conference on Computer Vision* (2017), 5842–5850.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *CoRR* abs/2006.07733 (2020). arXiv:2006.07733 https://arxiv.org/abs/2006.07733

[19] Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. NOVA - A tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 109–115. https://doi.org/10.1109/ACII.2019.8925519

[20] Lauren Klein, Victor Ardulov, Yuhua Hu, Mohammad Soleymani, Alma Gharib, Barbara Thompson, Pat Levitt, and Maja J. Matarić. 2020. Incorporating Measures of Intermodal Coordination in Automated Analysis of Infant-Mother Interaction. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 287–295. https://doi.org/10.1145/3382507.3418870

[21] C Leclère, M Avril, S Viaux-Savelon, N Bodeau, C Achard, S Missonnier, M Keren, R Feldman, M Chetouani, and D Cohen. 2016. Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3D reconstruction. *Translational Psychiatry* 6, 5 (May 2016), e816–e816. https://doi.org/10.1038/tp.2016.82

[22] Honggai Li, Jinshi Cui, Li Wang, and Hongbin Zha. 2020. Infant Attachment Prediction Using Vision and Audio Features in Mother-Infant Interaction. In *Pattern Recognition*, Shivakumara Palaiahnakote, Gabriella Sanniti di Baja, Liang Wang, and Wei Qi Yan (Eds.). Springer International Publishing, Cham, 489–502.

[23] Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2024. HEMM: Holistic Evaluation of Multimodal Foundation Models. arXiv:2407.03418 [cs.LG] https://arxiv.org/abs/2407.03418

[24] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR* abs/1906.08172 (2019). arXiv:1906.08172 http://arxiv.org/abs/1906.08172

[25] Ana Lyubenova, Dipika Neupane, Brooke Levis, Yin Wu, Yu Sun, Chunfeng He, Anjana Krishnan, Purnima Maharjan Bhandari, Zelalem Negeri, Muhammad Imran, Danielle B Rice, Marleine Azar, Matthew J Chiovitti, Nazanin Saadat, Kira E Riehm, Jill T Boruff, John P A Ioannidis, Pim Cuijpers, Simon Gilbody, Lorie A Kloda, Scott B Patten, Ian Shrier, Roy C Ziegelstein, Livia Comeau, Nicholas D Mitchell, Marcello Tonelli, Simone N Vigod, Franca Aceti, Jacqueline Barnes, Anupama Devi Bavle, Cheryl Tatano Beck, Carola Bindt, Philip M Boyce, Adomas Bunevicius, Linda H Chaudron, Nicolas Favez, Bárbara Figueiredo, Lluis Garcia-Esteve, Laura Giardinelli, Nadine Helle, Louise M Howard, Jane Kohlhoff, Lina Kusminskas, Zoltán Kozinszky, Lorenzo Lelli, Angeliki A Leonardou, Vasiliki Meuti, Sandra Nakić Radoš, Paloma Navarro García, Susan J Pawlby, Carlijn Quispel, Emma Robertson-Blackmore, Tamsen J Rochat, David J Sharp, Bon Wai M Siu, Alan Stein, Robert C Stewart, Meri Tadinac, S Darius Tandon, Iva Tendais, Annamária Töreki, Anna Torres-Giménez, Thach D Tran, Kylee Trevillion, Katherine Turner, Jose M Vega-Dienstmaier, Andrea Benedetti, and Brett D Thombs. 2021. Depression prevalence based on the Edinburgh Postnatal Depression Scale compared to Structured Clinical Interview for DSM Disorders classification: Systematic review and individual participant data meta-analysis. *International Journal of Methods in Psychiatric Research* 30, 1 (Mar 2021), e1860. https://doi.org/10.1002/mpr.1860

[26] Judi Mesman, Marinus H. van IJzendoorn, and Marian J. Bakermans-Kranenburg. 2009. The many faces of the Still-Face Paradigm: A review and meta-analysis. *Developmental Review* 29, 2 (2009), 120–162. https://doi.org/10.1016/j.dr.2009.02.001

[27] W. R. Mills-Koonce and M. Cox. 2013. Qualitative Ratings for Parent–Child Interaction at 3–48 Months of Age. (2013). Unpublished Manuscript.

[28] Mitho Müller, Ed Tronick, Anna-Lena Zietlow, Nora Nonnenmacher, Stephan Verschoor, and Birgit Träuble. 2016. Effects of Maternal Anxiety Disorders on Infant Self-Comforting Behaviors: The Role of Maternal Bonding, Infant Gender and Age. *Psychopathology* 49, 4 (09 2016), 295–304. https://doi.org/10.1159/000448404 arXiv:https://karger.com/psp/article-pdf/49/4/295/3492377/000448404.pdf

[29] Daniela Noe, Sarah Schluckwerder, and Corinna Reck. 2015. Influence of Dyadic Matching of Affect on Infant Self-Regulation. *Psychopathology* 48, 3 (2015), 173–183.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193 [cs.CV]

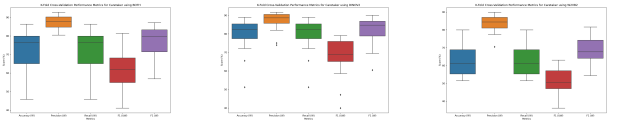
[31] Livio Provenzi, Renato Borgatti, Giorgia Menozzi, and Rosario Montirosso. 2015. A dynamic system analysis of dyadic flexibility and stability across the Face-to-Face Still-Face procedure: application of the State Space Grid. *Infant Behavior and Development* 38 (2015), 1–10. https://doi.org/10.1016/j.infbeh.2014.10.001
[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[33] Florian T. Ramseyer. 2020. Motion Energy Analysis (MEA). A primer on the assessment of motion from video. *Journal of Counseling Psychology* 67, 4 (2020), 536–549. https://doi.org/10.1037/cou0000407
[34] Claudia Reck, Daniela Noe, Francesca Cenciotti, Edward Tronick, and Karen M. Weinberg. 2009. *Infant and Caregiver Engagement Phases, German Revised Ed. (ICEP-R).* Unknown Publisher, Unknown Location.
[35] Corinna Reck, Alexandra Tietz, Miriam Müller, Katharina Seibold, and Edward Tronick. 2018. The impact of maternal anxiety disorder on mother-infant interaction in the postpartum period. *PLoS ONE* 13, 5 (2018), e0194763. https://doi.org/10.1371/journal.pone.0194763
[36] Dominik Schiller, Tobias Hallmen, Daksitha Withanage Don, Elisabeth André, and Tobias Baur. 2024. DISCOVER: A Data-driven Interactive System for Comprehensive Observation, Visualization, and ExploRation of Human Behaviour. arXiv:2407.13408 [cs.HC] https://arxiv.org/abs/2407.13408
[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402* (2012).
[38] Daniel Stern. 1971. A micro-analysis of mother-infant interaction: behaviors regulating social contact between a mother and her three-and-a-half-month-old twins. *Journal of the American Academy of Child Psychiatry* 10 (1971), 501–517. https://doi.org/10.1016/S0002-7138(09)61752-0
[39] Colwyn Trevarthen. 1979. Communication and Cooperation in Early Infancy: A Description of Primary Intersubjectivity. In *Before Speech: The Beginning of Human Communication*, Margaret Bullowa (Ed.). Cambridge University Press, London, 321–347.
[40] Colwyn Trevarthen and M. Bullowa. 1979. Communication and cooperation in early infancy: A description of primary intersubjectivity. *Before Speech (Cambridge)* (01 1979), 321–347.
[41] Edward Tronick, Heidelise Als, Lauren Adamson, Susan Wise, and T. Berry Brazelton. 1978. The Infant's Response to Entrapment between Contradictory Messages in Face-to-Face Interaction. *Journal of the American Academy of Child Psychiatry* 17, 1 (1978), 1–13. https://doi.org/10.1016/S0002-7138(09)62273-1
[42] Edward Tronick, Heidelise Als, Lauren Adamson, Stephen Wise, and T Berry Brazelton. 1978. The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child Psychiatry* 17, 1 (1978), 1–13.
[43] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. 2018. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access* 6 (2018), 1155–1166. https://doi.org/10.1109/ACCESS.2017.2778011
[44] Laurie A Van Egeren, Martha S Barratt, and Mark A Roach. 2001. Mother–infant responsiveness: Timing, mutual regulation, and interactional context. *Developmental Psychology* 37, 5 (2001), 684.
[45] Everett Waters. 1987. Attachment Q-Set (AQS).
[46] Marcia K. Weinberg and Edward Tronick. 1999. Infant and Caregiver Engagement Phases (ICEP): A behavioral system for assessing mutual regulation in infant-caregiver dyads. *Infant Mental Health Journal* 20, 1 (1999), 9–26.

A Appendix: Evaluation Method

To evaluate the performance of predictive models on the ICEP-R annotation task, we employed a two-pronged approach, starting with the exploration of a linear classification model and subsequently extending our analysis to a more complex bi-LSTM model.

A.1 Linear Classification Model

The architecture of the model comprises several fully connected layers: an initial layer with 512 units, followed by layers with 256 and 128 units, respectively. Each layer is followed by a GELU activation function and dropout regularization to prevent overfitting. The final output layer maps the 128-dimensional feature space to the number of target classes.



(a) Caregiver with Both Features **(b) Caregiver with DinoV2 Features** **(c) Caregiver with W2VB2 Features**

Figure 6: Evaluation of recall, precision, accuracy, and F1 (weighted and unweighted) for different feature sets.

A.2 bidirectional LSTM Architecture

The bi-LSTM classifier was trained and evaluated using a nested cross-validation approach combined with hyperparameter optimization. In this methodology, the dataset was first split into outer folds used for training and final evaluation, ensuring that the model's generalization capabilities were assessed across different subsets of the data. Within each outer fold, further cross-validation was performed on inner folds to fine-tune hyperparameters. This inner loop was critical for optimizing parameters such as window size, stride size, batch size, and the choice of optimizer, which included Adam and SGD, as well as learning rates.

The window size and stride size were particularly important, as they controlled the length of the input feature sequences and the degree of overlap between consecutive windows, respectively. These parameters directly influenced the model's ability to capture temporal context while balancing computational load. Batch size was another crucial factor, affecting both memory usage during training and the stability of gradient estimation. Finally, the choice of optimizer and learning rate significantly impacted the model's convergence and overall performance.

The outer folds were used to assess the model's generalization across various data subsets, ensuring that the model was not overfitting to any particular portion of the dataset. This rigorous approach provided a comprehensive evaluation of the model's performance.

A.3 Hyperparameter tuning for bi-LSTM

We employed a nested cross-validation strategy to evaluate the performance of different machine learning models for infant and caregiver mico-behavior, considering models with video-only, audio-only, and combined modalities. The outer cross-validation loop partitioned the dataset into training and testing subsets to ensure unbiased performance estimation. An inner cross-validation loop within each training subset optimized key hyperparameters such as window size, stride size, batch size, and optimizers.

During the hyperparameter tuning phase, smaller window sizes of 25 or 50 frames and strides of 25 or 50 were found to be more effective than larger sizes of 200 frames (~8 seconds long discreet annotation sampled at 25 fps). The smaller windows, each representing one second of data at a sampling rate of 25 frames per second, provided discrete, non-overlapping segments of data. This setup allowed the model to leverage its LSTM architecture to focus solely on immediate temporal features without redundancy from extensive historical data.

Moreover, optimal performance was observed with learning rates between 0.0001 and 0.00001, in conjunction with the SGD optimizer

Algorithm 1 Training and Evaluating the bi-LSTM Classifier

Define Hyperparameters
WINDOW ← Set of window sizes (e.g., [25, 50, 400])
STRIDE ← Set of stride sizes (e.g., [25, 50, 400])
BATCH_SIZES ← Set of batch sizes (e.g., [64, 128, 512])
OPTIMIZERS ← List of optimizers (e.g., ["Adam", "SGD"])
HIDDEN_SIZE ← set of lstm hidden layers (e.g., [128, 256, 512])
LEARNING_RATE ← Set of learning rates (e.g., [0.1e-4, 0.1e-10])
K_OUTER ← Number of folds (e.g., [10,20])
K_INNER ← List of optimizers (e.g., [5,10])
procedure SequenceData(Data, Window, Stride)
    Load and preprocess data
    Apply Sliding Window:
        Divide data into sequences using specified **Window** size
        Move the window over data with specified **Stride**
    Return grouped sequences with session names
end procedure
procedure NestedCrossValidation(Data, K_OUTER, K_INNER)
    for each outer fold **do**
        Split Data into TrainOuter and Test
        for each inner fold **do**
            Split TrainOuter into TrainInner and Validate
            Optimize hyperparameters on Validate
        end for
        Train model on TrainOuter
        Evaluate on Test
        Record performance metrics
    end for
end procedure

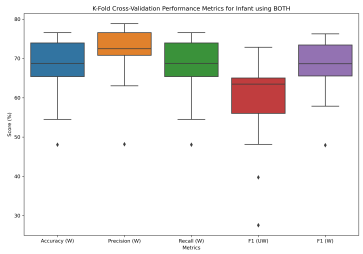
employing a momentum of 0.9. This learning rate range proved stable, aiding in consistent learning without fluctuations in loss.

Additionally, a weight decay (L1 regularization) of 0.1 effectively prevented early overfitting, thus maintaining model generalization and allowing the early stopping mechanism to activate under appropriate conditions. This configuration not only enhanced model performance but also optimized the learning process by balancing the exploration of temporal dynamics against the risk of overfitting on training data.

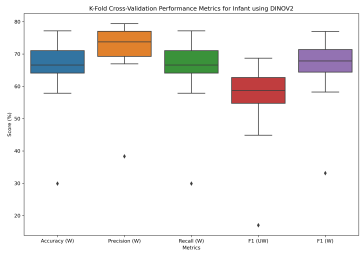
B Appendix: Results

B.1 bidirectional LSTM K-fold Results Box Plots

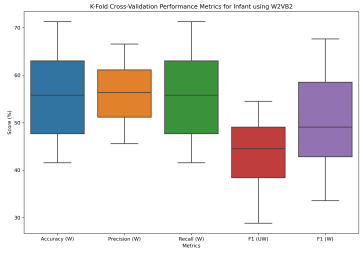
The infant multimodal model demonstrated consistent performance across all test folds, with a low standard deviation over 20 folds, achieving an average F1 score and accuracy of 68% with a 7 standard deviation in a 4-class prediction problem. Meanwhile, the caregiver dinov2 model achieved an average accuracy of 80% and a precision of 87% in a 3-class prediction task. We could observe a higher standard deviation due to the unbalance nature of the data.



(a) Infant with Both Features



(b) Infant with DinoV2 Features



(c) Infant with W2VB2 Features

Figure 7: Evaluation of recall, precision, accuracy, and F1 (weighted and unweighted) for different infant feature sets.