UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Engagement estimation in human-robot and human-human interactions across domain, age and culture: the MultiMediate ’26 challenge

Marius Funk, Daksitha Withanage Don, Michal Balazia, Victor Oei, Shogo Okada, François Brémond, Andreas Bulling, Elisabeth André, Philipp Müller

Angaben zur Veröffentlichung / Publication details:

Funk, Marius, Daksitha Withanage Don, Michal Balazia, Victor Oei, Shogo Okada, François Brémond, Andreas Bulling, Elisabeth André, and Philipp Müller. 2026. “Engagement estimation in human-robot and human-human interactions across domain, age and culture: the MultiMediate ’ 26 challenge.” In *MM’26: proceedings of the 34th ACM International Conference on Multimedia, Rio de Janeiro, Brazil, November 10-14, 2026*, edited by Debora Muchaluat-Saade, Laura Toni, Qin Jin, Vincent Oria, and Xavier Alameda-Pineda, 17836–41. New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3767308.3837718.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Engagement Estimation in Human-Robot and Human-Human Interactions across Domain, Age and Culture: The MultiMediate '26 Challenge

Marius Funk
University of Augsburg
Augsburg, Germany
JAIST
Nomi, Ishikawa, Japan
marius.funk@uni-a.de

Daksitha Withanage Don
University of Augsburg
Augsburg, Germany
daksitha.withanage.don@uni-a.de

Michal Balazia
INRIA Université Côte d'Azur
Sophia Antipolis, France
michal.balazia@inria.fr

Victor Oei
University of Stuttgart
Stuttgart, Germany
victor.oei@vis.uni-stuttgart.de

Shogo Okada
JAIST
Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp

François Brémond
INRIA Université Côte d'Azur
Sophia Antipolis, France
francois.bremond@inria.fr

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

Elisabeth André
University of Augsburg
Augsburg, Germany
elisabeth.andre@uni-a.de

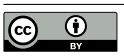
Philipp Müller
MPI for Intelligent Systems
Stuttgart, Germany
mueller@is.mpg.de

Abstract

Understanding engagement in human interaction is essential for modelling assistive and context-aware AI systems. Previously, the MultiMediate challenge has focused on estimating momentary continuous engagement values in adult conversation across cultures, achieving high levels of generalisability across domains. However, engagement can be measured in different metrics and manifest differently between children and adults. MultiMediate '26 therefore expands the evaluated domain set to include the PInSoRo dataset of Child-Child and Child-Robot play sessions in English. These sessions were annotated categorically with task and social engagement, diversifying engagement types, age ranges and interaction types. In this paper, we describe the MultiMediate '26 challenge and its goals, introduce the NoXi, NoXi-J, MPIIGI, and PInSoRo datasets, showcase extracted multi-modal features, present our baseline model for engagement estimation across all domains, and discuss challenge submissions and approaches. In an effort to provide a transparent and welcoming challenge environment, baseline code, leaderboard, datasets, and extracted features are publicly available at https://github.com/hcmlab/MultiMediate_26.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Computer vision*; **Supervised learning by regression**; **Supervised learning by classification**.



This work is licensed under a Creative Commons Attribution 4.0 International License.

MM '26, Rio de Janeiro, Brazil.

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3837718

Keywords

challenge; human-human interaction; human-robot interaction; dataset; engagement; domain adaptation

ACM Reference Format:
Marius Funk, Daksitha Withanage Don, Michal Balazia, Victor Oei, Shogo Okada, François Brémond, Andreas Bulling, Elisabeth André, and Philipp Müller. 2026. Engagement Estimation in Human-Robot and Human-Human Interactions across Domain, Age and Culture: The MultiMediate '26 Challenge. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3767308.3837718

1 Introduction

Estimating the momentary degree of user engagement is crucial for AI systems that take part in and support social interaction. Consequently, a growing body of work proposed approaches for the automatic estimation of engagement in human-human and human-machine interactions.

Engagement in human-human interactions was predicted using gaze features [5, 30] and multi-modal fusion approaches incorporating audio signals, gaze, eye contact, facial features and dialogue transcriptions [17, 18, 23, 44, 45, 47]. Human-agent and human-robot interactions were also analysed using backchannels [38], body posture [39] and multi-modal transformer models [19]. Research also showed success in predicting engagement in children [37], including interactions with artificial agents [14, 33].

MultiMediate is an expanding yearly challenge that focuses on human behaviour analysis in social interaction. It has addressed eye contact detection [28], backchannel detection [27], and backchannel recognition [26], while recent years have focused on the creation of engagement estimation approaches that are generalisable across cultures and conversational settings [25, 43]. While these efforts have led to significant advances in engagement estimation, they

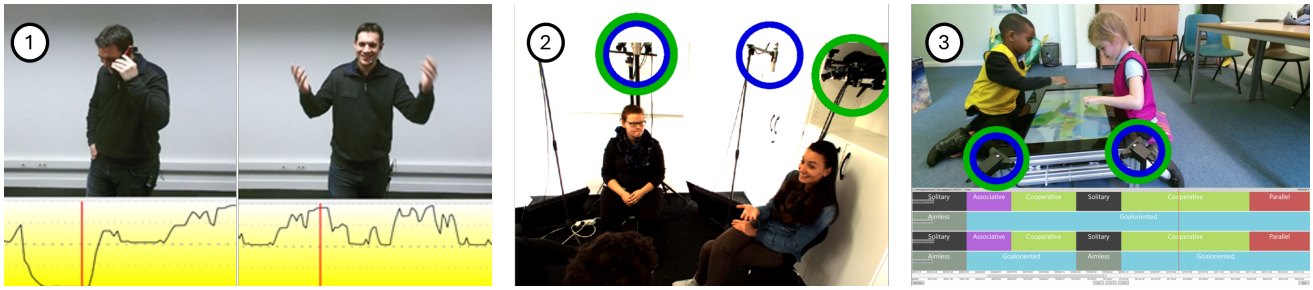


Figure 1: (1) Examples of low (left) and high (right) engagement in the NoXi corpus [6]. (2) Setup of the MPIIGroupInteraction dataset [29]. (3) Annotated Child-Child free-play session of the PInSoRo dataset [15] (© Séverin Lemaignan) displayed using the NOVA tool [4]. All images were used with the permission of the respective authors.

Table 1: MultiMediate '26 datasets: Languages and session count by dataset and split. *R* denotes continuous and *C* categorical engagement labels.

Dataset	Training Data	Validation Data	Test Data
NoXi (*R*) [6]	*English (23), French (7), German (8)*	*English (3), French (4), German (3)*	*English (6), French (6), German (4)*
NoXi Add. (*R*) [6]	-	-	*Indonesian (4), Spanish (4), Italian (2), Arabic (2)*
NoXi-J (*R*) [12]	*Japanese (21), Chinese (10)*	*Japanese (6), Chinese (4)*	*Japanese (6), Chinese (4)*
MPIIGI (*R*) [29]	-	*German (6)*	*German (6)*
PInSoRo CC (*C*) [15]	*English (20)*	*English (7)*	*English (6)*
PInSoRo CR (*C*) [15]	*English (12)*	*English (5)*	*English (6)*

remain constrained to a specific age group (adult), and a single notion of engagement (single overall magnitude per time point).

MultiMediate '26 significantly broadens the challenge scope by including the PInSoRo dataset [15, 16], a corpus of dyadic Child-Child and Child-Robot play sessions. These sessions differ from the established estimation task in significant ways. First, they show interactions of children with each other and with robots with only sparse adult participation. Second, they are not conceived as conversations, but as free play interactions with varying amounts of speech. Finally, engagement was not annotated using a continuous scale, but through two sets of categorical annotation labels: task engagement and social engagement.

2 Challenge Description

In MultiMediate '26, we focus on a multi-domain engagement estimation challenge across age groups, cultures, and engagement labels using a variety of datasets. Participating teams submit their predictions for the test set to an EvalAI server[1] for up to 20 submissions per team. We also continue our established challenge tasks of Bodily Behaviour Recognition, Backchannel Detection and Eye Contact Detection using the same server submission infrastructure.

[1]https://challenges.hcai.eu/web/challenges/challenge-page/29/overview

2.1 Task Definition

The task is to accurately predict different representations of engagement. We measure this using the Concordance Correlation Coefficient (CCC) [20], ranging from −1 to +1 for the datasets using continuous engagement annotation, namely NoXi, NoXi-J, NoXi (Add.), and MPIIGI. MultiMediate '26 additionally evaluates categorical social and task engagement for the PInSoRo dataset using Cohen's Kappa (κ), which also ranges from −1 to +1 and is therefore comparable to CCC. Using the average CCC score (CCC_{Avg}) and the average κ score (κ_{Avg}), the overall score is calculated as:

$$\textbf{Global} = \frac{2}{3}\,\text{CCC}_{\text{Avg}} + \frac{1}{3}\,\kappa_{\text{Avg}} \tag{1}$$

2.2 Datasets

MultiMediate '26 encompasses a wide variety of datasets and frame-wise engagement labels. The challenge contains recordings of interactions in nine languages for dyadic conversations, group interactions, and child play sessions including Child-Child and Child-Robot interactions. We continue the use of the *NO*vice e*X*pert *I*nteraction database (NoXi) [6] with its additional languages, NoXi (Add.), NoXi-J, the extension of NoXi recorded in Japan [12], and the MPIIGroupInteraction (MPIIGI) dataset [29]. An overview can be seen in Table 1. The datasets and all extracted features are available on the challenge website[2]. For the PInSoRo dataset, we share only extracted features.

This year, we extend our challenge with the Plymouth Interacting Social Robots (PInSoRo) dataset containing recordings of English-language free-play sessions in the form of Child-Child (20/7/6) and Child-Robot (12/5/6) interactions with an average duration of ~39 minutes [15, 16]. The unstructured interactions (see Figure 1, part 3) are dyadic with adult supervision, co-located around a touchscreen table, and were annotated categorically for social and task engagement [15]:

(1) *Task engagement.* The child's engagement in the free-play task divided into four categories: *goal-oriented*, *aimless*, *adult-seeking*, and *no play*.
(2) *Social engagement.* The child's social engagement divided into categories based on five of Parten's stages of play [34]: *solitary*, *onlooker*, *parallel*, *associative*, and *cooperative*.

[2]https://multimediate-challenge.org/Dataset/

All sessions of the validation and test sets are annotated by at least two raters. Besides annotations provided by Lemaignan et al. [15], ten additional CR sessions were annotated to guarantee this paradigm. We provide all available annotations of the train/validation set sessions. The test set is evaluated exclusively on the frames for each subtask in which all annotators agree on the label.

Training data policy. Participants can use the provided data in any configuration for model training (single multi-headed model, multiple models, domain adaptation). We ask participants to share their chosen approach for transparency. They are allowed to incorporate additional external open-source corpora during training.

3 Feature Extraction

Video, audio and text features were extracted from all challenge datasets and provided to the challenge participants. We do not share PInSoRo videos or transcripts due to privacy constraints.

3.1 Visual Features

Handcrafted Features: We used *OpenFace 2.0* [2] (714 Dim.) to extract gaze, facial landmarks and features, head pose, and action units as well as *OpenFace 3.0* [13] (21 Dim.) to extract eight action units, gaze direction and emotional labelling. *OpenPose* was used for 2-D pose estimation of hand and body keypoints [7] (139 Dim.). All features were extracted frame-wise.

Visual Embeddings: *DINOv2*'s ViT-B model [32] (768 Dim.), *Swin Transformer* [21, 22] (768 Dim.), and *VideoMAE v2* [42] (1408 Dim.) were used to extract visual embeddings with a stride of 16 frames. A frame-wise embedding space of joint text and image was derived via *CLIP* [35] (512 Dim.).

3.2 Audio and Text Features

Audio features include handcrafted and embedding-space auditive features as well as transcriptions and sentence embeddings. All features were extracted using the DISCOVER pipeline [40] with a 1 s window size and a 40 ms stride (except for the transcript). **NoXi** and **NoXi-J** audio is diarised by default as per the recording setup. Features for **MPIIGI** were extracted after human-annotated speaker diarisation. No diarisation was performed for PInSoRo audio features.

Audio features: We extracted a variety of frequency, energy and spectral parameters using the handcrafted extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS v2*) [11, 41] (88 Dim.). To complement the handcrafted features, we also provide the *w2v-BERT 2.0* model's embedding-space output [3] (1024 Dim.).

Text features: All speech was transcribed using *WhisperX* [1, 36]. Multilingual embeddings were then extracted from each transcribed utterance by *XLM-RoBERTa* [9] (768 Dim.).

4 Baseline Prediction

We use a feed-forward MLP with three ReLU hidden layers and dropout after the second hidden layer based on the MultiMediate '25 baseline architecture [43]. The hyperparameters were determined using KerasTuner Hyperband [31] with log uniform sampling over hidden widths, batch size, dropout rate and learning rate. All models are trained using the Adam optimiser.

Table 2: Test set baseline selection for *continuous* engagement labels.

Feature Set	NoXi	NoXi-J	MPIIGI	NoXi (Add.)	CCC_{Avg}
Video					
OpenFace 2.0	0.11	0.16	0.12	0.19	0.14
OpenFace 3.0	0.22	0.27	0.09	0.11	0.17
OpenPose	0.48	0.27	0.09	0.43	0.32
CLIP	0.44	0.22	-0.01	0.37	0.25
DINOv2	0.53	0.11	0.01	0.46	0.28
Swin Transformer	0.54	0.26	-0.05	0.46	0.31
VideoMAE	0.61	0.22	0.02	0.56	0.35
Audio					
eGeMAPS v2	**0.55**	**0.31**	**0.45**	**0.49**	**0.45**
w2v-BERT 2.0	0.60	0.32	0.05	0.47	0.36
Text					
XLM-RoBERTa	0.39	0.26	0.01	0.27	0.23

Table 3: Test set baseline selection for *categorical* engagement labels.

Feature set	CC Social	CC Task	CR Social	CR Task	κ_{Avg}
Video					
OpenFace 2.0	0.08	0.06	-0.01	0.25	0.09
OpenFace 3.0	0.13	0.10	0.02	0.17	0.10
OpenPose	0.06	0.03	-0.03	0.21	0.07
CLIP	0.09	0.12	-0.02	0.16	0.09
DINOv2	0.08	0.07	0.01	0.08	0.06
Swin Transformer	0.09	0.08	**0.09**	**0.24**	**0.13**
VideoMAE	0.14	0.13	-0.02	0.05	0.08
Audio					
eGeMAPS v2	0.13	0.02	-0.02	-0.03	0.03
w2v-BERT 2.0	0.11	0.05	0.00	0.03	0.05
Text					
XLM-RoBERTa	**0.14**	**0.16**	-0.03	0.10	0.10

4.1 Regression Models

The models estimating continuous labels were trained using a regression output layer and MSE loss with RMSE used as the performance metric. Test performance is reported as CCC. Based on the MultiMediate '25 baseline results, two models were trained on the NoXi and NoXi-J train and validation sets respectively. The model trained on NoXi was evaluated on the hold-out test sets for NoXi, NoXi (Add.), and MPIIGI. The NoXi-J test set was evaluated using the model trained on NoXi-J.

4.2 Categorical Models

Models estimating categorical labels use a multi-head approach for task and social engagement, the same trunk as the regression model, and a combined categorical cross-entropy loss. Per-sample weights were computed to account for class frequency. Two models were trained on the PInSoRo Child-Child (CC) and Child-Robot (CR) interactions train and validation sets and evaluated on their respective test set using Cohen's Kappa (κ) [8].

5 Results

5.1 Baseline Models

We trained models for each extracted feature on the train and validation splits of NoXi, NoXi-J, PInSoRo CC, and PInSoRo CR.

Table 4: Results of the multi-domain engagement estimation task.

Team	NoXi	NoXi-J	MPIIGI	NoXi (Add.)	CCC_{Avg}	CC Social	CC Task	CR Social	CR Task	κ_{Avg}	Global
Participants MM 2026											
HFUT-LMC	0.80	0.59	0.67	0.77	0.71	**0.65**	**0.48**	**0.46**	0.51	**0.52**	**0.65**
NUDT_LDM-AF	**0.82**	0.60	0.71	**0.80**	0.73	0.37	0.46	0.41	**0.58**	0.46	0.64
mltac	0.81	**0.66**	0.60	0.73	0.70	0.33	0.35	0.25	0.51	0.36	0.59
USTC_IAT_United	0.81	0.64	**0.75**	0.76	**0.74**	0.23	0.20	0.23	0.24	0.22	0.57
Winner MM 2025											
HFUT-LMC	0.79	0.58	0.67	0.75	0.70	–	–	–	–	–	–
Baseline (MM 2026)											
Baseline (ours)	0.55	0.31	0.45	0.49	0.45	0.14	0.16	0.09	0.24	0.16	0.36

Regression and cross-entropy model results are reported separately and then combined into the overall score. To include all datasets, we report the per-feature test scores. The chosen baseline should be seen as an upper bound on the introduced architecture.

Regression models. The overall best performing feature for the regression model was the *eGeMAPS v2* audio feature with an average CCC of 0.45 (see Table 2). It showed cross-domain generalisability with comparatively high performance for both in-domain (NoXi: 0.55, NoXi-J: 0.31) and out-of-domain (MPIIGI: 0.45, NoXi Add.: 0.49) test data.

Categorical models. The categorical models showed strong differences in performance between CC and CR interactions (see Table 3). While *XLM-RoBERTa* performed best for the CC subset with an average κ of 0.15 (CC Social: 0.14, CC Task: 0.16), the model underperformed for CR interactions (CR Social: -0.03, CR Task: 0.10). This mediocre performance was expected: in contrast to the lively CC sessions, most CR interactions consist of the child playing silently with the robot. *Swin Transformer* showed the best performance on the CR dataset (CR Social: 0.09, CR Task: 0.24), as well as the best overall performance with a κ of 0.13. The structural differences in interaction motivated us to choose the *XLM-RoBERTa* trained model for the CC baseline and the *Swin Transformer* model for the CR baseline.

Overall score. We combine the prediction results of *eGeMAPS v2* for the regression models (CCC_{Avg}: 0.45) and the average of the *XLM-RoBERTa* CC model performance and *Swin Transformer* CR model performance (κ_{Avg}: 0.16). We calculate the overall score with the formula defined in Eq. 1 (Global: 0.36).

5.2 Challenge Solution Results

We had 13 participating groups in the multi-domain engagement estimation challenge. The final leaderboard showing all published papers ranked by the global score can be seen in Table 4.

HFUT-LMC [46] won the challenge with a Semantic-Enhanced Sparse Temporal framework (SEST) that models engagement-related information from semantic and temporal perspectives using multimodal LLMs. They achieved a global score of 0.65 and the best average on the new PInSoRo dataset (0.52), performing best on all CC interaction tasks (Social: 0.65, Task: 0.48) and the CR social task (0.46). The model also outperforms the regression task score of DAPA (Domain-Adaptive Parallel Attention), their own domain prompting approach and last year's challenge winner [47].

The best performance on last year's tasks, however, was achieved by USTC_IAT_United [48] (0.74), who extended the DAPA model by structured multi-modal scaling and multi-party interaction generalisation, showing excellent performance in the few-shot adaptation to the MPIIGI dataset (0.75).

The most accurate NoXi (0.82) and NoXi (Add.) (0.80) predictions were achieved by NUDT_LDM-AF [10], who used Learnable Domain Memories and Adaptive Filtering (LDM-AF) to enable the model to efficiently learn from multiple domains. The best NoXi-J performance (0.66) was achieved by mltac [24] using Temporal Convolutional Networks and interaction modelling.

While these submissions outperformed the winner of last year's regression task, the progress seems to be plateauing, with NoXi only improving by 0.03 points. The newly introduced categorical prediction subtask exhibited considerable variation in model performance, indicating substantial potential for further advancement in future research.

The previous challenge tasks showed minor improvements: *Bodily Behaviour Recognition* accuracy improved from 0.65 to 0.67, *Eye Contact Detection* from 0.82 to 0.83. *Backchannel Detection* received one submission that did not improve accuracy.

6 Conclusion

MultiMediate '26 expanded the engagement estimation task by adding PInSoRo, English-language Child-Child and Child-Robot play sessions annotated with social and task engagement. Even though the additional subtask used classification instead of continuous annotation labels, several teams proposed new approaches or refined existing models that generalised knowledge across engagement label domains, improving upon last year's winning model and substantially outperforming the PInSoRo baselines.

Acknowledgments

Marius Funk was supported by a scholarship of the German Academic Exchange Service (DAAD). This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) through the Leibniz Award of Elisabeth André under Grant AN 559/10-1. Michal Balazia was funded by the French National Research Agency under the UCA^{JEDI} Investments into the Future, project number ANR-15-IDEX-01. This work has been partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG). We thank Séverin Lemaignan for his help with integrating the PInSoRo dataset in our challenge.

References

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023* (2023).

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 59–66. doi:10.1109/FG.2018.00019

[3] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187* (2023).

[4] 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). doi:10.1007/s13218-020-00632-3

[5] Roman Bednarik, Shahram Eivazi, and Michal Hradis. 2012. Gaze and Conversational Engagement in Multiparty Video Conversation: An Annotation Scheme and Classification of High and Low Levels of Engagement. In *Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. doi:10.1145/2401836.2401846

[6] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13 - 17, 2017*, Edward Lank, Alessandro Vinciarelli, Eve E. Hoggan, Sriram Subramanian, and Stephen A. Brewster (Eds.). ACM, 350–359. doi:10.1145/3136755.3136780

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299. doi:10.1109/CVPR.2017.143

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR* abs/1911.02116 (2019). arXiv:1911.02116 http://arxiv.org/abs/1911.02116

[10] Xuyao Deng, Qisheng Xu, Xie Chen, Yong Dou, Huaimin Wang, and Kele Xu. 2026. LDM-AF: Memory-Conditioned Adaptive Filtering for Cross-Domain Multimodal Engagement Estimation. In *Proceedings of the 34th ACM International Conference on Multimedia* (Rio de Janeiro, Brazil) *(MM '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3767308.3837720

[11] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202. doi:10.1109/TAFFC.2015.2457417

[12] Marius Funk, Shogo Okada, and Elisabeth André. 2024. Multilingual Dyadic Interaction Corpus NoXi+J: Toward Understanding Asian-European Non-verbal Cultural Characteristics and their Influences on Engagement. In *Proc. of the ACM International Conference on Multimodal Interaction*. 224–233. doi:10.1145/3678957.3685757

[13] Jiewen Hu, Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2025. OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis. *arXiv preprint arXiv:2506.02891* (2025).

[14] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić. 2020. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics* 5, 39 (2020). doi:10.1126/scirobotics.aaz3791

[15] Séverin Lemaignan, Charlotte ER Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PLOS one* 13, 10 (2018), e0205999.

[16] Séverin Lemaignan, James Kennedy, Paul Baxter, and Tony Belpaeme. 2016. Towards "machine-learnable" child-robot interactions: The PInSoRo dataset. In *Proceedings of the IEEE Ro-Man 2016 Workshop on Long-Term Child-Robot Interaction, New York, NY, USA*, Vol. 31.

[17] Jia Li, Yangchen Yu, Yin Chen, Yu Zhang, Peng Jia, Yunbo Xu, Ziqiang Li, Meng Wang, and Richang Hong. 2024. DAT: Dialogue-Aware Transformer with Modality-Group Fusion for Human Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. doi:10.1145/3664647.3688988

[18] Kun Li, Dan Guo, Guoliang Chen, Feiyang Liu, and Meng Wang. 2023. Data Augmentation for Human Behavior Analysis in Multi-Person Conversations. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9516–9520.

[19] Jia Yap Lim, John See, and Christian Dondrup. 2025. Multimodal Engagement Prediction in Human-Robot Interaction Using Transformer Neural Networks. In *MultiMedia Modeling*, Ichiro Ide, Ioannis Kompatsiaris, Changsheng Xu, Keiji Yanai, Wei-Ta Chu, Naoko Nitta, Michael Riegler, and Toshihiko Yamasaki (Eds.). Springer Nature Singapore, Singapore, 3–17.

[20] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268. doi:10.2307/2532051

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proc. of the IEEE/CVF International Conference on Computer Vision*. 10012–10022. doi:10.1109/ICCV48922.2021.00986

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3192–3201. doi:10.1109/CVPR52688.2022.00320

[23] Cheng Ma, Kevin Hyekang Joo, Alexandria K Vail, Sunreeta Bhattacharya, Álvaro Fernández García, Kailana Baker-Matsuoka, Sheryl Mathew, Lori L Holt, and Fernando De la Torre. 2025. Multimodal fusion with LLMs for engagement prediction in natural conversation. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction*. 244–259.

[24] Ane Cathrine Holst Merrild, Serkan Ayvaz, Stella Grasshof, Sami S. Brandt, and Martin Lund Trinhammer. 2026. Engagement is Tied to Context: A Configurable Multimodal Architecture for Engagement Estimation Across Adult and Child Interactions. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3840470.3841786

[25] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2024. MultiMediate'24: Multi-Domain Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. doi:10.1145/3664647.3689004

[26] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, et al. 2023. MultiMediate'23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9640–9645.

[27] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proc. of the ACM International Conference on Multimedia*. 7109–7114. doi:10.1145/3503161.3551589

[28] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proc. of the ACM International Conference on Multimedia*. 4878–4882. doi:10.1145/3474085.3479219

[29] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *Proc. of the ACM International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 153–164. doi:10.1145/3172944.3172969

[30] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 139–148.

[31] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. 2019. KerasTuner. https://github.com/keras-team/keras-tuner.

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=a68SUt6zFt Featured Certification.

[33] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education. In *Proc. of the AAAI Conference on Artificial Intelligence*. 687–694. doi:10.1609/aaai.v33i01.3301687

[34] Mildred B Parten. 1932. Social participation among pre-school children. *The journal of abnormal and social psychology* 27, 3 (1932), 243.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[37] Shyam Sundar Rajagopalan, O.V. Ramana Murthy, Roland Goecke, and Agata Rozga. 2015. Play with me — Measuring a child's engagement in a social interaction. In *Proc. of the IEEE International Conference and Workshops on Automatic*

Face and Gesture Recognition, Vol. 1. doi:10.1109/FG.2015.7163129

[38] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 375–382. doi:10.1109/HRI.2010.5453163

[39] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. In *Proc. of the ACM/IEEE International Conference on Human-robot Interaction*. 305–312.

[40] 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 preprint arXiv:2407.13408* (2024).

[41] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. of the International Workshop on Audio/Visual Emotion Challenge*. 3–10. doi:10.1145/2988257.2988258

[42] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14549–14560. doi:10.1109/CVPR52729.2023.01398

[43] Daksitha Senel Withanage Don, Marius Funk, Michal Balazia, Huajian Qiu, Shogo Okada, François Brémond, Jan Alexandersson, Andreas Bulling, Elisabeth André, and Philipp Müller. 2025. MultiMediate'25: Cross-cultural Multi-domain Engagement Estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14150–14155.

[44] Chunxi Yang, Kangzhong Wang, Peter Q Chen, MK Michael Cheung, Youqian Zhang, Eugene Yujun Fu, and Grace Ngai. 2023. MultiMediate 2023: Engagement Level Detection using Audio and Video Features. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9601–9605.

[45] Jun Yu, Xilong Lu, Lingsi Zhu, and Qiang Ling. 2025. LVLM-HBA: Large Vision-Language Model with Cross-Modal Alignment for Human Behavior Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery. doi:10.1145/3746027.3762077

[46] Yangchen Yu, Qian Chen, Jia Li, Zhenzhen Hu, Jinpeng Hu, Lizi Liao, Erik Cambria, and Richang Hong. 2026. Semantic-Enhanced Sparse Temporal Modeling for Dynamic Engagement Estimation in Multimodal Interactions. In *Proceedings of the 34th ACM International Conference on Multimedia* (Rio de Janeiro, Brazil) *(MM '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3767308.3837719

[47] Yangchen Yu, Yin Chen, Jia Li, Peng Jia, Yu Zhang, Li Dai, Zhenzhen Hu, Meng Wang, and Richang Hong. 2025. Generalizable Engagement Estimation in Conversation via Domain Prompting and Parallel Attention. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14170–14177.

[48] Yuefeng Zou, Guanjun Liu, Lingsi Zhu, Yichi Zhang, Zhichao Xia, and Jun Yu. 2026. UniEE: Heterogeneity-Aware Engagement Estimation across Multimodal Feature Representations, Interaction Structures, and Label Semantics. In *Proceedings of the 34th ACM International Conference on Multimedia*. Association for Computing Machinery. doi:10.1145/3767308.3837721