MultiMediate : Multimodal Behaviour Analysis for Artificial Mediation

Philipp Müller
mueller@is.mpg.de
Max Planck Institute for Intelligent Systems, Stuttgart
Germany

Marius Funk
Daksitha Withanage Don
Elisabeth André
Augsburg University
Germany

Victor Oei
Andreas Bulling
University of Stuttgart
Germany

Michal Balazia
François Brémond
INRIA Université Côte d'Azur, Sophia Antipolis
France

ABSTRACT

The MultiMediate challenge is a multi-year endeavour to lay the foundations for socially capable artificial mediators that can support human interactions. In the its first years, MultiMediate has focused on solving basic social behaviour sensing tasks including eye contact detection, backchannel detection, and bodily behaviour recognition. More recently, the challenge focused on the development of algorithms that can estimate the degree by which a human is engaged in a social interaction - a complex phenomenon that is strongly influenced by cultural norms and the nature of the social situation. By providing a diverse set of training and testing datasets, MultiMediate has facilitated the development of generalisable engagement estimation approaches. In MultiMediate '26, the diversity of training and evaluation data is enriched further by including the PInSoRo dataset of child-child and child-robot interactions annotated with both social and task engagement. As such, MultiMediate '26 poses the challenging task of creating engagement estimation approaches that are able to transfer between different social situations, languages, age groups, and notions of engagement.

1 INTRODUCTION

How social interaction unfolds has significant impacts on our private and professional lives. For example, if a shy person does not speak up during a brainstorming session, valuable ideas might be overlooked, and when teachers fail to engage students, learning success will not be optimal. One of the most ambitious, but also most promising, ways to support humans in social interaction is via an artificial mediator [44]. This interactive intelligent agent actively engages in social interaction in a human-like way to positively influence their course and/or outcomes. Among others, artificial mediators have been studied in mental health contexts [7], education [14, 32], or collaborative teamwork [8, 52]. While recent years have seen rapid development in language-based assistants [1, 45], the analysis and interpretation of fine-grained, multi-modal social behaviour still remains challenging [25, 54].

The goal of this multi-year challenge is to contribute to realising the vision of effective artificial mediators by measurable advances in central multi-modal social behaviour sensing and analysis tasks. The first iterations of MultiMediate ('21-'23) explored several important behaviour analysis tasks: eye contact detection, next speaker prediction, backchannel analysis, bodily behaviour recognition, and engagement estimation [35–37]. MultiMediate '24 and '25 focused on the development of engagement estimation approaches that generalize across domains. In MultiMediate '24, participants were provided with a main training set consisting of dyadic novice-expert interactions (NOXI; Cafaro et al. [9]) and were required to evaluate their engagement estimation approaches across a variety of different in-domain and out-of-domain test sets, which included different group compositions, interaction scenarios, or spoken languages. MultiMediate '25 added additional training data of dyadic novice-expert interactions between Japanese and Chinese speakers (NOXI-J; Funk et al. [16]). This increase in training data variety encouraged the development of dedicated approaches for cross-domain generalization (see e.g. Yu et al. [57]). Compared to previous iterations of the challenge this led to clear improvements in the ability to generalize across domains (Table 1).

In MultiMediate '26, we will further expand the range of domains covered in the engagement estimation challenge. In particular, we will incorporate the PInSoRo dataset of child-child and child-robot interactions annotated with engagement [26]. The PInSoRo dataset differs along several dimensions from the datasets already included in the MultiMediate challenge. First, the participants are from a different age group (children instead of adults). Second, it covers a new social situation (play instead of conversational situations). Third, in addition to human-human interactions, it also includes human-robot interactions. Finally, the structure of engagement annotations differs from the datasets included in previous iterations of MultiMediate . Instead of a single continuous engagement value, PInSoRo is annotated with nominal categorical values for social engagement and task engagement. Taken together, the inclusion of the PInSoRo dataset in the MultiMediate challenge will create a challenging test case for the generalization of existing engagement estimation approaches and will facilitate the development of novel methods able to accurately predict engagement across a wide variety of scenarios.

To ensure continuing progress, we will also invite submissions addressing the three most popular tasks of previous MultiMediate challenges: eye contact detection, backchannel detection, and bodily behaviour recognition. Future iterations of the challenge

Model	NOXI	NOXI-J	MPIGI	Additional	Combined
SOTA (MM 2025)					
DAPA (Yu et al [57])	**0.795**	**0.578**	**0.668**	**0.755**	**0.699**
Prior SOTA (MM 2024)					
DAT (Li et al [28])	0.760	–	0.490	0.670	–
Baseline (MM 2025)					
Baseline (ours)	0.570	0.440	0.130	0.470	0.400

Table 1: Results of the winning models for the multi-domain engagement estimation task in MultiMediate '25 and MultiMediate '24 (Metric: CCC, higher is better). Combined is the mean across NOXI, NOXI (Additional Languages), NOXI-J, and MPIIGroupInteraction. Inclusion of NoXi-J drastically improved generalizability and multi-domain performance (MPIGI, Additional).

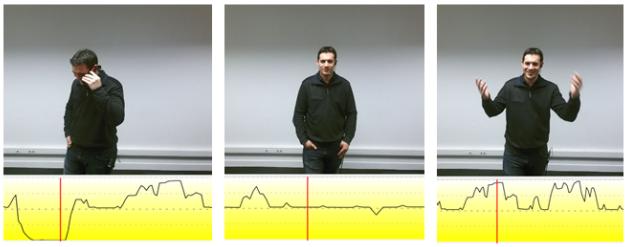

Figure 1: A participant in the NOXI corpus being disengaged (left), neutral (center) and highly engaged (right).

will continue to improve the generalisability of existing tasks and introduce novel, more complex and interactive tasks. These will include the generation of engaging agent behaviour, as well as social reasoning tasks.

2 RELATED WORK

Engagement has been investigated from various research angles, e.g. how to define, annotate, or automatically predict it. Several reviews indicate a growing interest in engagement in human-computer interaction [13, 18, 40]. Rich et al. [49] introduced a module for the recognition of engagement in human-robot interaction based on backchannels and Lim et al. [30] used multi-modal features in transformer models for engagement prediction. Sanghvi et al. [51] predicted engagement based on body posture features. Bednarik et al. [4] as well as Nakano and Ishii [39] focused on recognizing conversational engagement with gaze data. Inoue et al. [22] introduced an approach to engagement recognition using multimodal listener behaviours such as gaze, nodding and backchannels. Recently, Ma et al. [33] used a fusion model incorporating gaze, facial features and dialogue transcriptions for engagement prediction. Research in detecting engagement in students is prolific and promising [19, 24]. Engagement is also often studied in children [48] and, more particularly, in children interacting with an artificial agent [23, 41, 43]. Guhan et al. [20] researched engagement in mental health patients based on videos of the patient.

Some datasets also offer engagement ratings, such as MHHRI [11], RECOLA [50], UE-HRI [6], and a conversation dataset by Hradis, Bednarik, and Eivazi [5, 21]. Engagement estimation also was a task in MultiMediate '23 [35], leading to the development of several new multi-modal engagement estimation approaches [29, 55, 56].

What all approaches discussed above have in common is that they are trained on the same dataset on which their engagement estimation performance is evaluated. This is in stark contrast to common application scenarios, where training a model on the exact same data distribution that is encountered during testing is not feasible. To address this shortcoming, MultiMediate '24 [34] introduced the multi-domain engagement estimation challenge, which evaluated the ability of engagement estimation approaches to generalise across multiple domains. These different domains included conversations between speakers of different languages, as well as dyadic and group interactions. Participants' approaches were ranked by the average score achieved across the different domains. This approach was extended in MultiMediate '25 [54] by adding more sessions in additional languages (Japanese and Chinese) to the training set, significantly improving its generalizability and reducing the performance gap across domains as can be seen in the results of the engagement estimation challenge of MultiMediate '25 in Table 1.

3 ENGAGEMENT ESTIMATION CHALLENGE

3.1 Task Definition

Ground truth and evaluation metrics. The task includes the continuous, frame-wise prediction of conversational engagement of each participant in the social interaction. For the datasets NOXI, NOXI-J, NOXI (additional test languages), and MPIIGroupInteraction, engagement is annotated on a continuous scale from 0 (lowest) to 1 (highest). For PInSoRo, social and task engagement are annotated with a categorical scheme. In MultiMediate '26, we add to our task the PInSoRo dataset, which will contain categorical engagement values instead. Challenge participants are encouraged to investigate the multimodal as well as reciprocal behaviour of all recorded individuals. We will use the Concordance Correlation Coefficient (CCC) [31] to evaluate predictions on all test sets with continuous engagement annotations. For the categorical ground truth on PInSoRo, we will employ Cohen's Kappa. We choose Cohen's Kappa because it supports multi-class classification and in line with CCC it ranges from -1 to +1, with 0 indicating chance performance. This alignment between CCC and Cohen's Kappa is valuable when combining all test set results into a global score for the challenge.

Input features. We will provide a multi-modal set of pre-computed features to participants. From the audio signal, we provide transcripts generated with the Whisper model [47] and based on them sentence embeddings using XLM-RoBERTa [12]. In addition we supply GeMAPS features [15] along with wav2vec 2.0 embeddings [2]. From the video, we provide OpenFace [3] and OpenPose [10] outputs to cover facial as well as bodily behaviour. We also provide the vision features extracted by VideoMAEv2 [53], DINOv2 [42] and ImageBind [17] as well as contrastive language-image pretraining (CLIP) features [46]. These features will be provided on all training and test portions of the data.

Training Data	Validation Data	Test Data
NOXI [9] *English (23), French (7), German (8)*	NOXI [9] *English (3), French (4), German (3)*	NOXI [9] *English (6), French (6), German (4)*
		NOXI (additional test languages) [9] *Arabic (2), Italian (2), Indonesian (4), Spanish (4)*
	MPIIGroupInteraction [38] *German (6)*	MPIIGroupInteraction [38] *German (6)*
NOXI-J [16] *Japanese (21), Chinese (10)*	NOXI-J [16] *Japanese (6), Chinese (4)*	NOXI-J [16] *Japanese (6), Chinese (4)*
PInSoRo [26] *Child-Child (19), Child-Robot (15)*	PInSoRo [26] *Child-Child (5), Child-Robot (5)*	PInSoRo [26] *Child-Child (6), Child-Robot (6)*

Table 2: Engagement estimation datasets used in the MultiMediate '26 challenge. Languages and subsets covered by each dataset are given in italics, with the respective number of interactions in parentheses. The datasets without highlighting were already part of MultiMediate '25, the novel PInSoRo dataset containing only English interactions is highlighted in blue.

3.2 Training Datasets

We provide an overview of the different datasets used in MultiMediate '26 in Table 2. As training datasets, we provide NOXI and NOXI-J to our participants.

NOXI. The NOvice eXpert Interaction (NOXI) database [9] is a corpus of dyadic, screen-mediated face-to-face interactions in an expert-novice knowledge sharing context. In a session, one participant assumes the role of an expert and the other participant the role of a novice. Figure 1 shows two users during the interaction. NOXI includes interactions recorded at three locations (France, Germany and UK), spoken in seven languages (English, French, German, Spanish, Indonesian, Arabic and Italian), discussing a wide range of topics. The languages Indonesian, Arabic, Spanish, and Italian are not included in the training set of MultiMediate '26. They will serve as an out-of-domain evaluation set (described in Section 3.3). In total, the NOXI dataset offers over 25 hours (x2) of recordings of 84 dyadic interactions in natural settings, featuring synchronized audio, video (25fps), and motion capture data (using a Kinect 2.0). Following MultiMediate '23, we will use 48 interactions for training (comprising English, French and German recordings). For MultiMediate , each session was annotated in a continuous matter, meaning each video frame received a score between 0 and 1. Each rating was performed by at least two (up to 7) annotators (Average: 3.6 raters per session). We created gold standard annotations by calculating the mean over all raters. The NOXI dataset can be obtained from the website[1] after signing an EULA.

NOXI-J. The NOXI dataset was extended in 2024 by the NOXI-J dataset consisting of 66 dyadic interactions and over 16 hours of material recorded in Japan [16]. These additional interactions use the same setup as the original NOXI dataset. NOXI-J features 48 interactions in Japanese with native Japanese speakers and 18 interactions in Chinese with Chinese native speakers. In 34 of the Japanese interactions and all 18 Chinese interactions, participants gave their consent to sharing the video data with third-party researchers. As a result, we included a total of 52 new interactions in MultiMediate '25. Of these 52 new interactions, we provide 31 as training data (21 Japanese, 10 Chinese), and 10 as validation data (6 Japanese, 4 Chinese). Additionally to the training interactions used for MultiMediate '24, we will include 42 sessions of the NOXI-J dataset for an overall test set of 90 dyadic conversations. Engagement annotations for NOXI-J were created in the same manner as for the NOXI dataset using a minimum of 3 (up to 5) annotators. Similarly to NOXI, the dataset can be obtained from our website[2].

PInSoRo. For MultiMediate '26, we extend our challenge by the Plymouth Interacting Social Robots (PInSoRo)[3] dataset containing recordings of free-play English-language sessions with 45 Child-Child and 30 Child-Robot interactions with overall 120 children and a total duration of 45 hours and 48 minutes [26, 27]. The interactions were annotated categorically for social- and task engagement [26]:

(1) *Task engagement.* Annotation of the child's engagements in the free-play task divided into four categories: goal oriented, aimless, adult seeking, and no play.
(2) *Social engagement.* Annotation of the social interaction during play between participants divided into five categories: solitary, onlooker, parallel, associative, and cooperative.

In MultiMediate '26, we will provide 34 interactions as training data (19 Child-Child, 15 Child-Robot) and 10 as validation data (5 Child-Child, 5 Child-Robot) for an additional categorical prediction task that aims to further improve generalizability of our engagement estimation approach. For the prediction targets, we will provide all available annotations of the train set sessions and include annotated frames in which all annotators agree on the label for each measurement in the validation set.

3.3 Evaluation Datasets

We make use of four different evaluation datasets to quantify performance across different domains. For the MPIIGroupInteraction data set, feature modalities differ from NOXI (additional group members). For participants to better adapt to this dataset, we provide a

[1] https://multimediate-challenge.org/datasets/Dataset_NoXi/

[2] https://multimediate-challenge.org/datasets/Dataset_NoXiJ/

[3] https://freeplay-sandbox.github.io/dataset

labelled validation set that may be used for evaluation and even for training supervised domain adaptation approaches.

NOXI (MultiMediate '23 version). This part of the evaluation set is identical to the test set of MultiMediate '23 and consists of 16 sessions (in English, French and German). That is, the MultiMediate '23 version of the NOXI test set comes from the same domain as the training set, providing a reference to compare MultiMediate '26 submissions to MultiMediate '23-'25 results, as well as a point of comparison for evaluating the impact of out-of-domain test scenarios on performance.

NOXI (additional test languages). This evaluation set includes four languages that are not part of the NOXI training set: two sessions in Arabic, two in Italian, four in Indonesian, and four in Spanish. As a result, this evaluation set tests the ability of participants' approaches to transfer to new languages and cultural backgrounds not seen at training time.

MPIIGroupInteraction. Following previous iterations of MultiMediate , we make use of the MPIIGroupInteraction corpus[4], consisting of 22 group discussions between three to four people, each lasting for 20 minutes [38]. The MPIIGroupInteraction dataset has the distinct advantage of the availability of six unpublished group discussions that can be used for evaluation as a test set. For MultiMediate '23 we collected novel engagement annotations on the MPIIGroupInteraction test and validation sets. The validation set with ground truth annotations will be provided to participants to monitor their performance on the out-of-domain task. In addition, it may be used as a limited set of training data to develop supervised domain adaptation approaches.

NOXI-J. This evaluation set was included in MultiMediate '25 and adds 6 Japanese and 4 Chinese language sessions from the same domain as the NOXI-J training set. This set will show how the training of a more culturally varied prediction model affects its performance.

PInSoRo. For MultiMediate '26, we newly include this evaluation set that adds PInSoRo interactions from the same domain as the PInSoRo training and validation sets. It contains 6 Child-Child and 6 Child-Robot interactions and only includes frame-wise annotations that are agreed-upon by at least two annotators for each of the three annotation measurements.

3.4 Continuing Tasks

To facilitate continuing progress, we will include the three most popular tasks from earlier iterations of MultiMediate as independent tracks in MultiMediate '26: eye contact detection, backchannel detection, and bodily behaviour recognition. Past iterations of MultiMediate have already shown the effectiveness of this approach. E.g., while the winning approach for eye contact detection in MultiMediate '21 reached an accuracy of 0.56 on the official test set, the winner of MultiMediate '25 was able to reach 0.82 accuracy [5].

[4]https://multimediate-challenge.org/datasets/Dataset_MPII/
[5]https://multimediate-challenge.org/leaderboards/leaderboard_eyecontact/

4 EVALUATION APPROACH

To compare the submissions to the challenge in a fair and coherent way, we will make the test datasets (without ground truth) available to participants two weeks before the challenge deadline. Participants, in turn, will submit their predictions for evaluation against the ground truth on our servers. Participants' approaches will be ranked on a leaderboard using the average performance (CCC) across all test datasets. In line with previous years of MultiMediate we intend to invite solution paper submissions from challenge participants. These paper submissions will be evaluated according to the performance achieved and their scientific quality.

5 FUTURE AGENDA

Our future agenda consists of three main directions: (1) Further improving generalization, (2) introducing interactive tasks, and (3) addressing more open ended social reasoning problems.

In order to improve generalization, we will continue to record, curate, and annotate suitable social interaction datasets. Concretely, we have already annotated the training set of the MPIIGroupInteraction dataset with engagement scores. This additional training data is ready to be integrated into the engagement challenge in MultiMediate '27. We are also actively exploring opportunities to establish multi-domain evaluations for other tasks such as bodily behaviour recognition, or backchannel detection.

At present, we are developing a framework to conduct interactive experiments to investigate the generation of engaging (or disengaging) agent behaviour. In the context of MultiMediate , we plan to pose the challenge of creating agent behaviour that is engaging according to different cultural norms. The NOXI and NOXI-J datasets will be highly useful for this purpose as they contain examples of highly engaging and non-engaging explanations for different cultural scenarios. As evaluation of generated behaviour is not straightforward, we plan to investigate a combination of evaluation approaches such as using large language models and humans via services such as Mechanical Turk. We anticipate to include this task in either MultiMediate '27 or MultiMediate '28.

To generate richer and more diverse training data for artificial mediators, we plan to collect fine-grained social reasoning annotations on the MultiMediate datasets. These will consist of diverse questions such as *What did person B laugh about?* or *Why did person A interrupt person C?*, with associated multiple-choice answers. This approach will be helpful in training artificial mediators that can engage more deeply in human interactions. We expect the dataset and associated task to be available after MultiMediate '28.

6 STATEMENT OF COMMITMENT AND CONTACT INFORMATION

To realise the future agenda of the MultiMediate challenge, we will publish and maintain a website for at least three years. We will provide the appropriate information, datasets and tasks and will rigorously evaluate the submissions. We are looking forward to working with the organisers of the ACM Multimedia Conference to publicise the challenge tasks and invite researchers for participation. The organisers who will be responsible for organising, publicising, reviewing and judging the Grand Challenge submissions are listed

in the following. Philipp Müller is the primary contact author.

Dr. Philipp Müller
Max Planck Institute for Intelligent Systems
Stuttgart, Germany
mueller@is.mpg.de

Prof. Dr. Elisabeth André
Augsburg University
Augsburg, Germany
andre@informatik.uni-augsburg.de

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

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

ACKNOWLEDGMENTS

Marius Funk was supported by a scholarship of the German Academic Exchange Service (DAAD).

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 59–66.

[4] 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 *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction* (Santa Monica, California) *(Gaze-In '12)*. ACM, New York, NY, USA, Article 10, 6 pages.

[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 *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction (Gaze-In '12)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/2401836.2401846

[6] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[7] Chris Birmingham, Zijian Hu, Kartik Mahajan, Eli Reber, and Maja J. Mataric. 2020. Can I Trust You? A User Study of Robot Mediation of a Support Group. *arXiv preprint arXiv:2002.04671* (2020).

[8] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8.

[9] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar. 2017. The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In *International Conference on Multimodal Interaction*. ACM.

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

[11] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2019. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing* 10, 4 (Oct. 2019), 484–497. https://doi.org/10.1109/TAFFC.2017.2737019 Conference Name: IEEE Transactions on Affective Computing.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[13] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* 51, 5, Article 99 (nov 2018), 39 pages. https://doi.org/10.1145/3234149

[14] Olov Engwall and José Lopes. 2020. Interaction and collaboration in robot-assisted language learning for adults. *Computer Assisted Language Learning* (2020), 1–37.

[15] 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.

[16] 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 *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 224–233. https://doi.org/10.1145/3678957.3685757

[17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

[18] Nadine Glas and Catherine Pelachaud. 2015. Definitions of Engagement in Human-Agent Interaction. 944–949. https://doi.org/10.1109/ACII.2015.7344688

[19] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2021. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review* 33, 1 (March 2021), 27–49. https://doi.org/10.1007/s10648-019-09514-z

[20] Pooja Guhan, Naman Awasthi, and Kathryn McDonald, Kristin Bussell, Dinesh Manocha, Gloria Reeves, and Aniket Bera. 2022. MET: Multimodal Perception of Engagement for Telehealth. http://arxiv.org/abs/2011.08690 arXiv:2011.08690.

[21] Michal Hradis, Shahram Eivazi, and Roman Bednarik. 2012. Voice activity detection from gaze in video mediated communication. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, Santa Barbara California, 329–332. https://doi.org/10.1145/2168556.2168628

[22] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Latent Character Model for Engagement Recognition Based on Multimodal Behaviors. In *9th International Workshop on Spoken Dialogue System Technology*, Luis Fernando D'Haro, Rafael E. Banchs, and Haizhou Li (Eds.). Springer, Singapore, 119–130.

[23] 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). https://doi.org/10.1126/scirobotics.aaz3791

[24] Shofiyati Nur Karimah and Shinobu Hasegawa. 2021. Automatic Engagement Recognition for Distance Learning Systems: A Literature Study of Engagement Datasets and Methods. In *Augmented Cognition (Lecture Notes in Computer Science)*, Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.). Springer International Publishing, Cham, 264–276. https://doi.org/10.1007/978-3-030-78114-9_19

[25] Dong Won Lee, Yubin Kim, Denison Guvenoz, Sooyeon Jeong, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. 2025. The Human Robot Social Interaction (HSRI) Dataset: Benchmarking Foundational Models' Social Reasoning. *arXiv preprint arXiv:2504.13898* (2025).

[26] 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.

[27] 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.

[28] 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)*. https://doi.org/10.1145/3664647.3688988

[29] 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.

[30] 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.

[31] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268.

[32] José Lopes, Olov Engwall, and Gabriel Skantze. 2017. A first visit to the robot language café. In *ISCA workshop on Speech and Language Technology in Education*.

[33] 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.

[34] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, et al. 2024. MultiMediate'24: Multi-Domain Engagement Estimation. In *32nd ACM International Conference on Multimedia*. 11377–11382.

[35] 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.

[36] 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 *Proceedings of the 30th ACM International Conference on Multimedia*. 7109–7114.

[37] 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 *Proceedings of the 29th ACM International Conference on Multimedia*. 4878–4882. https://doi.org/10.1145/3474085.3479219

[38] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, Tokyo, Japan, 153–164. https://doi.org/10.1145/3172944.3172969

[39] 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.

[40] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI* 7 (Aug. 2020). https://doi.org/10.3389/frobt.2020.00092

[41] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI* 7 (2020). https://www.frontiersin.org/articles/10.3389/frobt.2020.00092

[42] 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.

[43] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 687–694. https://doi.org/10.1609/aaai.v33i01.3301687

[44] Sunjeong Park and Youn-kyung Lim. 2020. Investigating User Expectations on the Roles of Family-shared AI Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[45] Soya Park, Hari Subramonyam, and Chinmay Kulkarni. 2023. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering. *arXiv preprint arXiv:2312.06024* (2023).

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[47] 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.

[48] 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 *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1. 1–8. https://doi.org/10.1109/FG.2015.7163129

[49] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 375–382.

[50] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. 1–8. https://doi.org/10.1109/FG.2013.6553805

[51] 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 *Proceedings of the 6th International Conference on Human-robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. ACM, New York, NY, USA, 305–312.

[52] Elaine Short and Maja J. Mataric. 2017. Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 385–390.

[53] 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. https://doi.org/10.1109/CVPR52729.2023.01398

[54] 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.

[55] 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.

[56] Jun Yu, Keda Lu, Ji Zhao, Zhihong Wei, Iek-Heng Chu, and Peng Chang. 2024. Dialogue cross-enhanced central engagement attention model for real-time Engagement estimation. In *33rd Int'l Joint Conf. on Artificial Intelligence*. 3187–3195.

[57] 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.