MultiMediate '24: Multi-Domain Engagement Estimation

Philipp Müller
DFKI
Saarbrücken, Germany
philipp.mueller@dfki.de

Michal Balazia
INRIA Sophia Antipolis
Sophia Antipolis, France
michal.balazia@inria.fr

Tobias Baur
University of Augsburg
Augsburg, Germany
tobias.baur@uni-a.de

Michael Dietz
University of Augsburg
Augsburg, Germany
michael.dietz@uni-a.de

Alexander Heimerl
University of Augsburg
Augsburg, Germany
alexander.heimerl@uni-a.de

Anna Penzkofer
University of Stuttgart
Stuttgart, Germany
anna.penzkofer@vis.uni-stuttgart.de

Dominik Schiller
University of Augsburg
Augsburg, Germany
dominik.schiller@uni-a.de

François Brémond
INRIA Sophia Antipolis
Sophia Antipolis, France
francois.bremond@inria.fr

Jan Alexandersson
DFKI
Saarbrücken, Germany
janal@dfki.de

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

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

Abstract

Estimating the momentary level of participant's engagement is an important prerequisite for assistive systems that support human interactions. Previous work has addressed this task in within-domain evaluation scenarios, i.e. training and testing on the same dataset. This is in contrast to real-life scenarios where domain shifts between training and testing data frequently occur. With MultiMediate '24, we present the first challenge addressing multi-domain engagement estimation. As training data, we utilise the NOXI database of dyadic novice-expert interactions. In addition to within-domain test data, we add two new test domains. First, we introduce recordings following the NOXI protocol but covering languages that are not present in the NOXI training data. Second, we collected novel engagement annotations on the MPIIGroupInteraction dataset which consists of group discussions between three to four people. In this way, MultiMediate '24 evaluates the ability of approaches to generalise across factors such as language and cultural background, group size, task, and screen-mediated vs. face-to-face interaction. This paper describes the MultiMediate '24 challenge and presents baseline results. In addition, we discuss selected challenge solutions.

CCS Concepts

• **Human-centered computing**; • **Computing methodologies** → **Artificial intelligence**;

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3689004

Keywords

challenge, dataset, engagement, nonverbal behaviour, domain adaptation

ACM Reference Format:
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 (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3664647.3689004

1 Introduction

Knowing how engaged humans are in a conversation is an important prerequisite for many assistive systems, in particular if their aim is to maintain a high level of participation. As a result, the estimation of human engagement has become an active research field, addressing a wide variety of approaches and scenarios. These cover engagement prediction in human-human [13, 21], and human-agent interactions [14, 28, 32], as well as for different age groups including adults [12, 21], students [11, 15], or children [26, 31]. There is a wide variety of features that are utilised for engagement estimation, including backchannels [32], pose features [33], or gaze data [6]. The inclusion of engagement estimation in the MultiMediate '23 challenge lead to several new multi-modal engagement estimation approaches [13, 35, 37, 38].

What all these approaches have in common is that they are trained and evaluated on the same dataset each. While Guhan et al. [12] applied their model trained on the MEDICA dataset on separately collected data, they did not evaluate engagement estimation performance on this separate dataset. Despite clear progress on the engagement estimation task, within-domain testing does not reflect domain shifts that are frequent when applying approaches in the

real world. The complex nature of engagement makes it prone to influences of context variables. Engagement might be expressed differently by people of different cultures, in different group compositions (dyadic vs. more than two people), and might be subject to different task characteristics.

In MultiMediate '24 we pose the challenge of creating engagement estimation approaches that are able to transfer across such context factors. To this end we significantly extend the engagement estimation task of MultiMediate '23 by two new out-of-domain evaluation sets. In particular, we introduce a not-yet released multilingual variant of the NOXI corpus [7] to cover a wide variety of additional languages and cultural backgrounds, including Indonesian, Arabic, Spanish, and Italian. Furthermore, we employ novel engagement annotations on the MPIIGroupInteraction corpus [25], which consists of groups of three to four people engaged in face-to-face discussions. Taken together, these evaluation sets vary along several dimensions: language and cultural background, group size, task, and screen-mediated vs. face-to-face interaction. MultiMediate '24 is embedded in a multi-year challenge with the goal of addressing several nonverbal behaviour analysis tasks relevant to autonomous artificial mediators. The first iteration of the challenge in 2021 [23] has addressed eye contact detection and next speaker prediction while MultiMediate '22 focused on backchannel analysis [1, 22]. MultiMediate '23 addressed bodily behaviour recognition [3, 21] and engagement estimation on the NOXI corpus [7].

In this paper, we define the multi-domain engagement estimation task, the evaluation criteria, and describe new annotations collected on the NOvice eXpert Interaction (NOXI) database [7], as well as on test and validation portions of MPIIGroupInteraction [25]. Furthermore, we present baseline approaches for the challenge task and report evaluation results. We make all collected annotations, baseline implementations, and raw feature representations publicly available for further use, also beyond the scope of MultiMediate '24.[1]

2 Challenge Description

In the following we present challenge task and the utilised training and testing datasets. Testing data (without ground truth) was released to participants before the challenge deadline. Participants in turn submitted their predictions for evaluation.

2.1 Task definition

In line with MultiMediate'23 [21], the engagement estimation task addresses the frame-wise prediction of the conversational engagement level of each interlocutor on a continuous scale from 0 (lowest) to 1 (highest). To evaluate predictions on the test datasets, we make use of the Concordance Correlation Coefficient (CCC) [19] which ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). The key difference to the engagement task in MultiMediate '23 is that MultiMediate '24 poses the challenge to address engagement estimation in a multi-domain evaluation scenario. For this multi-domain evaluation, we employ two new out-of-domain test datasets: NOXI (Additional Languages) and the MPIIGroupInteraction test set. Challenge participants are encouraged to develop methods that can generalise across these different domains and make use of multi-modal as well as reciprocal behaviour of both interlocutors.

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

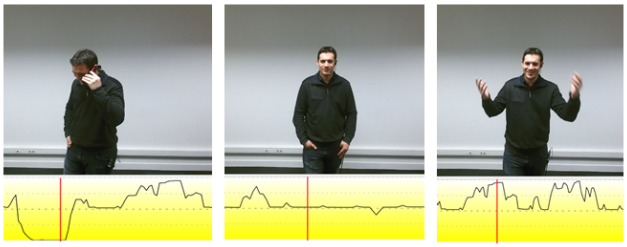

Figure 1: A participant in the NOXI corpus being disengaged (left), neutral (center) and highly engaged (right).

2.2 Datasets

For the multi-domain engagement estimation task, we utilise three different datasets. The training portion of the NOXI corpus [7] is used as the main training dataset. Testing is performed on three different test sets. First, the in-domain NOXI test set which was already utilised in MultiMediate'23. Second, an out-of-domain version of NOXI that includes conversations in languages not present in the NOXI training set. Finally, we utilise test and validation portions of the MPIIGroupInteraction dataset [25] that were annotated with engagement labels specifically for this challenge. In the following, we will present each dataset in detail.

NOXI. For training, we follow MultiMediate '23 and make use of the NOvice eXpert Interaction (NOXI) database [7, 21]. NOXI is a corpus of dyadic, screen-mediated interactions in an expert-novice knowledge sharing context. In each session, one participant assumed the role of an expert and the other participant the role of a novice. Figure 1 shows a user during interaction. The goal of NOXI was to obtain data of spontaneous behaviour in a natural setting on a variety of discussion topics. Therefore, one of the main design goals was to match recorded participants based on their common interests. In a first step, potential experts were gathered who expressed their willingness to share their knowledge about one or more topics they were knowledgeable and passionate about. In a second step, novices were recruited that were willing to discuss or learn more about the available set of topics offered by experts. The recording protocol furthermore introduced interruptions of the novices in order to provoke experts' reactions when conversational engagement gets interrupted. NOXI includes interactions recorded at three locations (France, Germany and UK), spoken in eight languages (English, French, German, Spanish, Indonesian, Arabic, Dutch and Italian), discussing a wide range of topics. The dataset offers over 25 hours (x2) of interaction recordings, featuring synchronized audio, video (25fps), and motion capture data (using a Kinect 2.0). For training, and within-domain testing, we use a subset of this corpus containing 48 sessions for training and 16 sessions for testing (75/25 split). These sessions cover the languages English, French, German, and Dutch. For MultiMediate '23, each session was annotated in a continuous matter, meaning each video frame has a score between 0 and 1. Each rating was performed by

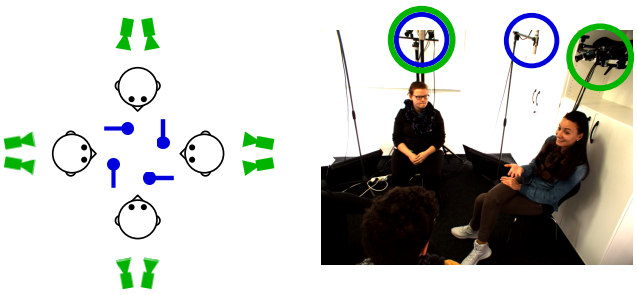

Figure 2: Setup of the MPIIGroupInteraction dataset. Reproduced with permission from the authors of [25].

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[2].

NOXI (Additional 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. We annotated these interactions for MultiMediate '24, following the same protocal as in MultiMediate '23 [21].

MPIIGroupInteraction. To test the performance of participant's approaches in a different social situation, we make use of the MPIIGroupInteraction corpus[3] [25]. This corpus consists of audiovisual recordings of group discussions between three to four participants, each lasting for 20 minutes. MPIIGroupInteraction differs from NOXI in several key aspects. First, it consists of group discussions while NOXI features dyadic interactions. Second, there are no predefined roles such as "novice" or "expert", only the task to discuss on a topic that was selected to be controversial among the group members. Third, interactions are face-to-face instead of screen-mediated. Finally, participants are seated throughout the whole interaction. With these differences to the NOXI setup, MPIIGroupInteraction presents a challenging evaluation case for engagement estimation approaches. For MultiMediate '23 we collected novel engagement annotations on the MPIIGroupInteraction test and validation sets. The validation set with ground truth annotations is 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. The validation set comprises 6 recordings with 21 participants, while the test set consists of 6 recordings with 23 participants.

3 Experiments

We first present the different features we extracted on all datasets. These features were used for our baseline experiments, and were also given to participants to develop their challenge solutions.

[2]https://multimediate-challenge.org/datasets/Dataset_NoXi/
[3]https://multimediate-challenge.org/datasets/Dataset_MPII/

3.1 Visual Features

On both the NOXI and the MPIIGroupInteraction dataset, participants' locations or seating positions are known. As a result, we can directly extract visual features, without the need to first localise and track participants.

Head Features. We extracted features from participants' head and face using OpenFace 2.0 [4]. All features where extracted for each video frame. The resulting feature vectors are consisting of 68 3D facial landmarks, 56 3D eye landmarks, presence and intensity of 18 action units as well as markers for detection success, detection certainty, facial position and rotation.

Pose Features. We extract body pose estimates using OpenPose [8], resulting in a 139-dimensional feature representation including 2D body, hand, and facial keypoints for every video frame.

CLIP Embeddings. As a general visual representation, we extract CLIP (Contrastive Language-Image Pretraining) [29] embeddings for each video frame. CLIP is trained to learn a joint embedding space for text and images. In this way, it can capture a wide variety of semantic content present in the videos, e.g. relating to emotional expressions, attention, and many more. The clip feature embeddings have 512 dimensions.

3.2 Audio Features

To analyse human speech for behavioural insights, it is essential to differentiate between its verbal and vocal components. The verbal aspect pertains to the use of words and language to express ideas, thoughts, or information. It encompasses the content of what is said. In contrast, the vocal component relates to the sounds produced by the voice, including tone, pitch, volume, and other characteristics of how something is said. Both verbal and vocal features have been extracted using the DISCOVER framework [34].

Vocal Features. For the paralinguistic assessment of engagement, we extracted two sets of features using a one-second sliding window with a 40 ms stride, aligned with the video stream's frame rate. The first set is the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [10], which includes 54 acoustic parameters frequently used in tasks such as depression, mood, and emotion recognition [36]. The second set of features was obtained using a pretrained w2v-BERT 2.0 encoder [5], which provides automatically learned representations of the audio signal. This model was trained unsupervised on a large dataset comprising 4.5 million hours of audio and has shown exceptional performance in various downstream tasks, including speech-to-text and expressive speech-to-speech translation.

Verbal Features. To analyze the verbal content of spoken language, it is essential to first convert speech into text (STT). STT systems have been a focal point of research for many years. For our STT module, we utilize WhisperX [2], an adaptation of the Whisper Model [30], which offers enhanced timestamp accuracy, support for longer audio sequences, and faster transcription. When extracting features from text, the language of the text plays a crucial role. Since the challenge datasets are recorded in multiple languages, we employ multilingual textual feature extraction using the XLM RoBERTa model developed by Conneau et al. [9]. This model is

	NOXI		NOXI (Add. Languages)		MPIIGroupInteraction		Combined
Features	Val CCC	Test CCC	Val CCC	Test CCC	Val CCC	Test CCC	Test CCC
Video							
OpenFace 2.0	0.81	0.28	-	0.13	**0.09**	0.00	0.14
OpenPose	0.83	0.48	-	0.41	0.01	0.06	0.32
CLIP	**0.88**	0.48	-	0.38	-0.01	0.06	0.31
Voice							
eGemaps v2	0.77	0.56	-	0.47	0.00	**0.15**	0.39
w2vbert2	0.77	**0.64**	-	**0.51**	0.05	0.09	**0.41**
Text							
XLM RoBERTa	0.62	0.40	-	0.29	0.00	0.00	0.23

Table 1: Concordance correlation coefficient (CCC) of different featuresets on different engagement estimation validation and test sets. For NOXI (Additional Languages) no validation set is available.

based on a transformer architecture and has been trained on a vast collection of multilingual data from the internet. To preserve the semantics of the transcript, every speech segment overlapping with the sliding window is included.

3.3 Baseline Prediction Approach

We evaluated the utility of the different feature modalities presented above for the task of frame-wise engagement estimation. We implemented a fully connected neural network consists of an input layer followed by three hidden layers of size 136 each. To prevent overfitting we relied on a dropout layer after the second hidden layer with a dropout rate of 0.25. The network was trained using the Adam optimizer and the mean squared error loss function. All hyperparameters were optimized using the hyperband search algorithm of the KerasTuner framework [27]. We trained all approaches on the NOXI training set and evaluated on the three different test sets. In particular, we did not use the MPIIGroupInteraction validation set for training. Our baseline implementation is available online[4].

4 Results

We first discuss the results of our baseline experiments and then give an overview over the results achieved by teams participating in the challenge.

4.1 Baseline Experiments

The baseline results are depicted in Table 1. The best average performance across all test sets is achieved by w2vbert2 features with an average CCC of 0.41, followed by eGemaps v2 features (0.39 average CCC). Video features lack behind with OpenPose reaching the best performance at 0.32 average CCC. Text features from XLM RoBERTa only achieve 0.23 average CCC. All featuresets suffer from domain shifts. This is especially severe when testing on the MPIIGroupInteraction dataset, where even the best approach (eGemaps v2) only achieved 0.15 CCC. The impact of the domain shift on NOXI (Additional Languages) is less severe. Nevertheless, all featuresets are impacted when comparing to the standard NOXI test set results. For voice and text features, the degradation ranges from 0.09 CCC (eGempas v2) to 0.13 CCC (w2vbert2). For visual features the impact has a similar range, from 0.07 CCC (OpenPose) to 0.15 CCC (OpenFace 2.0). The fact that visual features are impaired indicates that in addition to speaking another language, there is also a shift in how visual nonverbal behaviour expresses engagement, pointing to the impact of different cultural norms that pose a challenge to generalisation.

4.2 Challenge Solution Results

We provide the challenge leaderboard for the multi-domain engagement estimation task in Table 2. Approaches are ranked by the average test error across all test datasets. In total, 10 approaches were able to surpass the baseline. The best approach by the team USTC-IAT-United was able to reach an average CCC of 0.68, representing an increase over the baseline by 0.27 CCC. Two papers describing challenge solutions were accepted for publication at ACM Multimedia [16, 17], both focusing on the multi-modal fusion mechanisms. Li et al. [17] proposed to first fuse modality-specific representations across interactants and subsequently perform modality fusion. This approach reached an average CCC of 0.64 and also set a new state-of-the-art on the in-domain NOXI test set with 0.76, outperforming TCA-NET which previously reached 0.75 CCC on the NOXI test set [13]. Kumar et al. [16] in contrast proposed an approach which only processes features obtained from a single target participant. They investigated different strategies to fuse modalities in a hierarchical fashion, reaching 0.64 average CCC in the challenge. With respect to the cross-domain generalization abilities of participants' approaches, the gap between the original NOXI test set and NOXI (Additional Languages) is usually small, and in some cases non-existent (e.g. UST-IAT-United). The approach of Li et al. [17] showed the largest gap, but also reached the highest performance on the original NOXI test set, indicating a larger degree of overfitting. The domain gap between NOXI and MPIIGroupInteraction is still much larger. Future approaches could investigate dedicated domain adaptation protocols to close this gap.

In addition to the multi-domain engagement estimation task, we also invited submissions to selected tasks from previous iterations of MultiMediate . An especially noteworthy method was proposed by Ma et al. [20] who reached a new state of the art on the

[4]https://git.opendfki.de/philipp.mueller/multimediate24

Rank	Approach	NOXI	NOXI (Add. Languages)	MPIIGroupInteraction	Combined
1	USTC-IAT-United	0.72	**0.73**	**0.59**	**0.68**
2	AI-lab	0.69	0.72	0.54	0.65
3	Li et al. [17]	**0.76**	0.67	0.49	0.64
4	Kumar et al. [16]	0.72	0.69	0.50	0.64
5	ashk	0.72	0.69	0.42	0.61
6	YKK	0.68	0.66	0.40	0.58
7	Xpace	0.70	0.70	0.34	0.58
8	nox	0.68	0.70	0.31	0.56
9	SP-team	0.68	0.65	0.34	0.56
10	YLYJ	0.60	0.52	0.30	0.47
11	Baseline (ours)	0.64	0.51	0.09	0.41

Table 2: Challenge leaderboard for the engagement estimation task. Team names are replaced by references in case of accepted publications.

eye contact detection challenge [23, 24]. The authors proposed an adaptive feature selection method which can reduce computational burden while reaching high prediction accuracy. They obtained an accuracy of 0.79, improving over the previous state of the art at 0.777 accuracy [18].

5 Conclusion

We introduced MultiMediate '24, the first challenge addressing engagement estimation in a multi-domain evaluation scenario. We presented novel annotations on publicly available datasets, including pre-computed feature representations. Furthermore, we defined the evaluation protocol, presented baseline results, and discussed successful challenge solutions. Overall, we observed that while transfer between the original NOXI dataset and NOXI (Additional Languages) tends to work well, a larger domain gap remains when testing on MPIIGroupInteraction. Datasets and evaluation will be accessible to researchers even beyond the MultiMediate challenge, contributing to continuing progress on the challenge tasks.

Acknowledgments

P. Müller and J. Alexandersson were funded partially by the European Union Horizon Europe programme, grant number 101078950. A. Bulling was funded by the European Research Council (ERC; grant agreement 801708). A. Penzkofer was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 39074001. M. 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 was supported in part by the Deutsche Forschungsgemeinschaft (DFG) through the Leibniz Award of E. André under Grant AN 559/10-1.

References

[1] Ahmed Amer, Chirag Bhuvaneshwara, Gowtham K Addluri, Mohammed M Shaik, Vedant Bonde, and Philipp Müller. 2023. Backchannel Detection and Agreement Estimation from Video with Transformer Networks. *arXiv preprint arXiv:2306.01656* (2023).

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

[3] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. 2022. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proc. of the ACM International Conference on Multimedia*. 70–79. https://doi.org/10.1145/3503161.3548363

[4] 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. https://doi.org/10.1109/FG.2018.00019

[5] 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).

[6] 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*. https://doi.org/10.1145/2401836.2401846

[7] 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 *Proc. of the International Conference on Multimodal Interaction*. https://doi.org/10.1145/3136755.3136780

[8] 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. https://doi.org/10.1109/CVPR.2017.143

[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] 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. https://doi.org/10.1109/TAFFC.2015.2457417

[11] 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 (2021), 27–49. https://doi.org/10.1007/s10648-019-09514-z

[12] Pooja Guhan, Naman Awasthi, Kristin Bussell, Dinesh Manocha, Gloria Reeves, Aniket Bera, et al. 2020. Developing an Effective and Automated Patient Engagement Estimator for Telehealth: A Machine Learning Approach. *arXiv preprint arXiv:2011.08690* (2020).

[13] Hongyuan He, Daming Wang, Md Rakibul Hasan, Tom Gedeon, and Md Zakir Hossain. 2024. TCA-NET: Triplet Concatenated-Attentional Network For Multimodal Engagement Estimation. In *Proceedings of the IEEE International Conference on Image Processing*.

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

[15] 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)*. Springer International Publishing, Cham, 264–276. https://doi.org/10.1007/978-3-030-78114-9_19

[16] Deepak Kumar, Surbhi Madan, Pradeep Singh, Abhinav Dhall, and Balasubramanian Raman. 2024. Towards Engagement Prediction: A Cross-Modality Dual-Pipeline Approach using Visual and Audio Features. In *Proceedings of the 32nd ACM International Conference on Multimedia*. https://doi.org/10.1145/3664647.3688986

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

[20] Fuyan Ma, Yiran He, Bin Sun, and Shutao Li. 2024. Less is More: Adaptive Feature Selection and Fusion for Eye Contact Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. https://doi.org/10.1145/3664647.3688987

[21] 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.

[22] 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. https://doi.org/10.1145/3503161.3551589

[23] 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. https://doi.org/10.1145/3474085.3479219

[24] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proc. of the ACM Symposium on Eye Tracking Research & Applications*. 1–10. https://doi.org/10.1145/3204493.3204549

[25] 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. https://doi.org/10.1145/3172944.3172969

[26] 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://doi.org/10.3389/frobt.2020.00092

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

[28] 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. https://doi.org/10.1609/aaai.v33i01.3301687

[29] 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.

[30] 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.

[31] 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. https://doi.org/10.1109/FG.2015.7163129

[32] 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. https://doi.org/10.1109/HRI.2010.5453163

[33] 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.

[34] 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).

[35] Vu Ngoc Tu, Van Thong Huynh, Hyung-Jeong Yang, Soo-Hyung Kim, Shah Nawaz, Karthik Nandakumar, and M Zaigham Zaheer. 2023. DCTM: Dilated Convolutional Transformer Model for Multimodal Engagement Estimation in Conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9521–9525.

[36] 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. https://doi.org/10.1145/2988257.2988258

[37] 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.

[38] Jun Yu, Keda Lu, Mohan Jing, Ziqi Liang, Bingyuan Zhang, Jianqing Sun, and Jiaen Liang. 2023. Sliding Window Seq2seq Modeling for Engagement Estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9496–9500.