MultiMediate '25: Cross-cultural Multi-domain Engagement Estimation

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

Marius Funk
University of Augsburg
Augsburg, Germany
marius.funk@uni-a.de

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

Huajian Qiu
University of Stuttgart
Stuttgart, Germany
huajian.qiu@vis.uni-stuttgart.de

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

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

Jan Alexandersson
DFKI
Saarbrücken, Germany
janal@dfki.de

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
DFKI
Saarbrücken, Germany
philipp.mueller@dfki.de

Abstract

Estimating momentary conversational engagement is central to assistive, socially aware AI systems, yet models are typically trained and evaluated within a single domain, limiting real-world robustness. The MultiMediate '25 challenge advances engagement estimation to more challenging, cross-cultural, and multi-domain settings. Building on prior challenge editions, we expand beyond NOXI as the sole training source by introducing NOXI-J, a new multilingual corpus covering Japanese and Chinese interactions, enabling both training and evaluation in diverse linguistic contexts. Although NOXI-J conceptually extends NOXI, we treat it as a distinct domain because linguistic, cultural, capture, and annotation differences induce measurable distribution shifts. In this paper, we present new annotations, precomputed multi-modal features (visual, vocal, and verbal), baseline evaluations, and an analysis of the best performing challenge solutions. Beyond accuracy, we quantify fairness using *Conditional Demographic Disparity* for gender and language. Our baselines confirm strong in-domain performance (e.g., paralinguistic eGeMAPS and video-transformer features) and reveal notable cross-domain drops, underscoring the challenge of cultural, linguistic, and interactional shifts. Fairness analyses indicate generally small discrepancies for our baselines. We observe the largest disparities for the proposed challenge solutions on the Chinese language test set. All annotations, features, code, and leaderboards are made publicly available to foster sustained progress on robust and fair engagement estimation.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762076

CCS Concepts

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

Keywords

challenge, dataset, engagement, domain adaptation

ACM Reference Format:
Daksitha 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 (MM '25), Oct. 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746027.3762076

1 Introduction

Knowing how engaged humans are in a conversation is a prerequisite for many assistive and mediation systems, especially when the goal is to sustain participation and improve interaction quality. Consequently, *engagement estimation* has become an active research area across human–human [13, 22] and human–agent interactions [14, 30, 34], spanning adults [12, 22], students [11, 15], and children or infants [27, 33, 41]. Methodologically, prior work has leveraged diverse behavioral cues, e.g., conversational backchannels [34], gross body pose and motion [35], gaze [5], and paralinguistics [9].

The MultiMediate challenge has progressively expanded the behavioral analysis agenda from low-level behaviours to more complex social phenomena. Utilising the MPIIGroupInteraction dataset, MultiMediate '21 targeted eye contact detection and next-speaker prediction [24, 25], MultiMediate '22 focused on backchannel detection and agreement estimation [23], and MultiMediate '23 introduced bodily behaviour recognition [2, 22]. In addition, MultiMediate '23 introduced the engagement estimation task on the NOXI corpus, which is a multilingual corpus of dyadic, screen-mediated novice–expert conversations, providing synchronized audio-video recordings for analyzing social signals and engagement. [6]. To explicitly probe generalization, MultiMediate '24 defined a *multi-domain* evaluation training on NOXI and testing also on (i) NOXI recordings in additional languages and (ii) MPIIGroupInteraction (3–4 person face-to-face discussions) revealing notable cross-domain drops [21].

However, prior engagement editions provided *training* data only in three European languages (English, French, German), with additional languages appearing solely as evaluation sets (e.g., Arabic, Italian, Indonesian, Spanish) [21]. This limited the community's ability to study cross-cultural generalization. In this year's edition, we address this limitation by introducing new training corpora in Japanese and Chinese through the NOXI-J extension [10], and by making both NOXI and NOXI-J available for model development. Though NOXI-J extends NOXI, we treat them as separate domains: linguistic, cultural, capture, and annotation differences induce distribution shift, enabling clean cross-domain tests and reproducible per-domain reporting.

This year, we go beyond accuracy by reporting Conditional Demographic Disparity (CDD) with respect to *gender* and *language*, aligning with European non-discrimination principles [38]. Motivating this focus, socio-linguistic work shows that conversational cues (e.g., prosodic triggers for backchannels) vary across languages [40], while upstream speech technology can exhibit demographic performance gaps that propagate downstream [16].

To quantify both within-corpus accuracy and cross-domain transfer, we report two baseline tracks—one trained on NOXI and one on NOXI-J each evaluated on NOXI, NOXI-J, NOXI (Additional Languages), and MPIIGroupInteraction [21]. This separation isolates the benefit of newly available Japanese and Chinese training data while preserving comparability to prior editions. All annotations, features, and baseline code are released to facilitate future work beyond MultiMediate '25.[1]

2 Challenge Description

MultiMediate '25 poses a *cross-cultural, multi-domain* engagement estimation challenge. The evaluation spans speakers of Japanese, Chinese, German, Arabic, Indonesian, and French, and covers both dyadic and multi-party interactions. Test are released without ground truth and teams submit frame-wise predictions for evaluation on a EvalAI server [2]. In addition to engagement, we continue to welcome submissions to established tasks from prior editions: eye contact detection, bodily behaviour recognition, and backchannel detection.

[1]https://multimediate-challenge.org
[2]https://challenges.hcai.eu/

2.1 Task definition

The task is *frame-wise* prediction of each interlocutor's engagement on a continuous scale $[0, 1]$. Accuracy is measured with the Concordance Correlation Coefficient (CCC), ranging from -1 to $+1$. Participants are free to use the provided labelled data for training and validation and undergo in-domain and out-of-domain evaluations on NoXI, NoXI-J, NoXI (Additional Languages), and MPIIGroupInteraction.

Training data policy. Two training corpora are provided: NOXI and NOXI-J. We *do not prescribe* usage. Participants may (i) train a single model on the union, (ii) train separate models and select/fuse at test time, or (iii) apply domain adaptation using only the provided train/val splits. In addition, few-shot adaptation on the MPIIGroupInteraction *validation* set is permitted for out-of-domain tuning (not the test set). For transparency, teams should report their choice. Our baselines include both a NOXI-trained and a NOXI-J-trained model to bracket these strategies.

Fairness evaluation (CDD). We quantify group-wise bias with CDD, which measures average prediction differences at the same ground-truth level y; conditioning on y asks whether, for equal true engagement, the model systematically over- or under-predicts for a group. We report CDD_G for gender. We acknowledge that gender is not binary; however, in our datasets only male/female self-identifications are available.

$$\mathrm{CDD}_G \;=\; \mathbb{E}[\hat{y} \mid \text{male},\, y] \;-\; \mathbb{E}[\hat{y} \mid \text{female},\, y]\,,$$

where positive values indicate higher predictions for males than females at the same y, and negative values the reverse. Furthermore, we define CDD_L to measure CDD for each language.

$$\mathrm{CDD}_L \;=\; \mathbb{E}[\hat{y} \mid L,\, y] \;-\; \mathbb{E}[\hat{y} \mid y]\,,$$

where language L is compared to the pooled expectation across all languages at the same y.

Since y is continuous, we approximate the conditionals by binning y (e.g., 10 equal-frequency bins) and averaging group differences within bins, weighted by bin prevalence. Values closer to 0 indicate smaller disparities.

2.2 Datasets

We evaluate cross-domain, cross-cultural generalization using dyadic and multi-party corpora spanning Japanese (JA), Chinese (ZH), German (DE), Arabic (AR), Indonesian (ID), French (FR), and English (EN) as illustrated in Table 1. All datasets are available online.[3]

NOXI (train/val/test). NOXI [6] is a corpus of screen-mediated *dyadic* expert–novice interactions with synchronized audio/video and frame-wise engagement annotations ranging from 0 (lowest engagement) to 1 (highest engagement). These annotations were first collected forMultiMediate '23 [22]. Following MultiMediate '23 and MultiMediate '24, we use EN/FR/DE for train/val/test.

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

Table 1: Engagement estimation datasets used in MultiMediate '25. Languages per split are shown in italics with number of interactions in parentheses.

Training Data	Validation Data	Test Data
NOXI [6] *English (23), French (7), German (8)*	NOXI [6] *English (3), French (4), German (3)*	NOXI [6] *English (6), French (6), German (4)* NOXI (Additional Languages) [6] *Arabic (2), Italian (2), Indonesian (4), Spanish (4)*
	MPIIGroupInteraction [26] *German (6)*	MPIIGroupInteraction [26] *German (6)*
NOXI-J [10] *Japanese (21), Chinese (10)*	NOXI-J [10] *Japanese (6), Chinese (4)*	NOXI-J [10] *Japanese (6), Chinese (4)*

NOXI (Additional Languages) (test only). An out-of-domain NOXI split comprising AR/IT/ID/ES (test only) probes cross-language transfer beyond the EN/FR/DE training languages. Annotations follow the same protocol as NOXI.

MPIIGroupInteraction (val/test). MPIIGroupInteraction contains *group* discussions (3–4 participants, ∼20 minutes) recorded face-to-face [26]. It differs from NOXI in interaction format (multi-party vs. dyadic), roles (no expert/novice), and setting (co-located vs. screen-mediated). We provide a validation split (6 recordings, 21 participants) with labels and a test split (6 recordings, 23 participants) without labels for evaluation.

NOXI-J (train/val/test). NOXI-J extends NOXI with Japanese and Chinese *dyadic* sessions recorded with the same setup [10]. For this challenge we release JA (21 train, 6 val, 6 test) and ZH (10 train, 4 val, 4 test) with frame-wise engagement from ≥ 3 raters per session (labels are rater means). This enables training on Asian languages in addition to European ones.

3 Experiments

We extract audio, visual, and text features on all corpora and release them to participants together with our baseline code.

3.1 Visual Features

On NOXI and MPIIGroupInteraction, speaker locations/seating are known, so we can extract per-person visual features without an additional tracking stage. In particular, we provide the following features. *Head/Face features from (OpenFace 2.0),* including 3D facial and eye landmarks (68+56), action unit presence/intensity (18), and quality/pose indicators for each frame [3]. *Body Pose (OpenPose)*, consisting of 2D body, hand, and facial keypoints, yielding a 139-D representation per frame [7]. *CLIP embeddings*, 512 dimensions per frame [31]. *DINOv2 embeddings* [29]: we derive per-frame visual tokens and apply PCA to obtain a compact representation (reduced to 768×3); features are sampled every 16 frames (stride 16). *Video Backbones (Swin; VideoMAE v2):* We compute 768-D Video-Swin embeddings [19, 20] and 1408-D VideoMAE v2 embeddings [39] on non-overlapping 16-frame clips (stride 16), capturing local spatiotemporal context.

3.2 Audio Features

We separate *vocal* (paralinguistic) and *verbal* (content) streams; both are extracted with the DISCOVER pipeline [36]. To represent vocal information, we compute eGeMAPS features using a 1 s window with 40 ms hop [9, 37], as well as self-supervised w2v-BERT 2.0 embeddings [4]. We represent verbal information by transcribing speech with WhisperX [1, 32]. Multilingual sentence embeddings are obtained with XLM-RoBERTa [8], aggregating all transcript segments overlapping each analysis window.

3.3 Baseline Prediction Approach

We cast frame-wise engagement estimation as a regression problem. The baseline is a feed-forward MLP (three ReLU hidden layers; linear output) with dropout after the second hidden layer. Models are trained with Adam and MSE; RMSE is monitored during training. Hyperparameters are selected via KerasTuner Hyperband [28] with compact ranges: hidden widths $u_1, u_2, u_3 \in [8, 512]$ (step 8); dropout $d \in [0, 0.5]$ (step 0.05); learning rate $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}$; batch size $b \in \{32, 64, \ldots, 2048\}$; batch size is tuned jointly with the other hyperparameters as part of the search.

Our primary performance metric is the Concordance Correlation Coefficient (CCC) [18]. Fairness is quantified with Conditional Demographic Disparity (CDD) [38] for gender (CDD_G) and language (CDD_L); values are centered at 0 (no disparity) and the sign indicates direction.

We evaluate baselines trained on either NoXI or NOXI-J ib fizr gekd-out test sets: NOXI, NOXI-J, NoXI (Additional Languages), and MPIIGroupInteraction. For each baseline we report a *Baseline CCC*, the unweighted mean CCC across its four test sets, which we use to rank feature modalities. Reference code (Hyperband search, CCC/CDD) is available.[4]

4 Results

4.1 Baseline Experiments

We report two kinds of baselines: one trained on NOXI (train+val) and one on NOXI-J (train+val). Each baseline method is evaluated on four held-out test sets, NOXI, NOXI-J, NOXI (Additional Languages), and MPIIGroupInteraction. Furthermore, we computed *CCC* defined as the unweighted mean CCC across its four test sets; this scalar ranks modalities and serves as the leaderboard reference.

For the baselines trained on NOXI (Table 2), eGeMAPS v2 is the most robust across domains (Combined CCC of 0.40), combining strong in-domain accuracy on NOXI (0.57) with solid transfer to NOXI-Additional (0.47) and MPI (0.44). Among video encoders, VideoMAE leads (0.36 Combined CCC), followed by Swin (0.28). Text-only XLM-RoBERTa lags (0.20). For the baselines trained on NOXI-J (Table 3), VideoMAE attains the best Combined CCC (0.14),

[4] https://git.opendfki.de/philipp.mueller/multimediate25

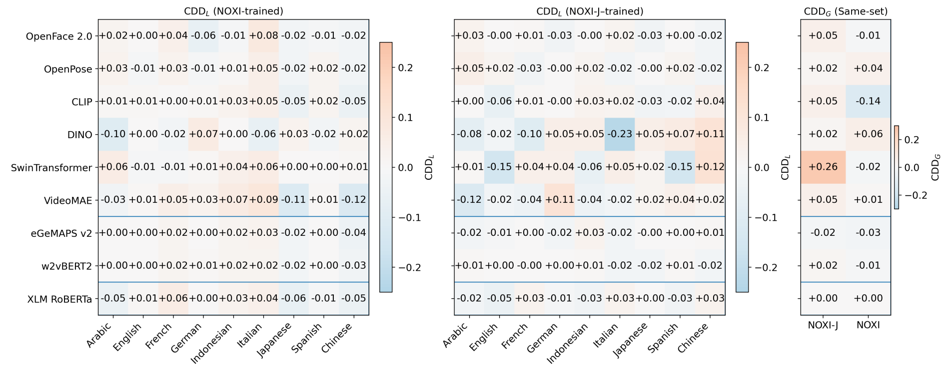

Figure 1: Fairness heatmaps for the baseline models. Left: CDD$_L$ per language for the NOXI-trained baseline (evaluated on NOXI, NOXI (Additional), and NOXI-J). Middle: CDD$_L$ per language for the NOXI-J–trained baseline (same evaluations). Right: CDD$_G$ (male minus female) for models trained and tested on the same corpus.

Table 2: CCC values across test datasets for baseline models trained on NOXI.

Feature Set	NOXI	NOXI-J	MPIGI	Additional	Combined
Video					
OpenFace 2.0	0.22	0.00	0.01	0.09	0.08
OpenPose	0.47	0.03	0.03	0.39	0.23
CLIP	0.48	0.04	0.00	0.38	0.23
DINO	0.29	**0.14**	0.08	0.06	0.14
SwinTransformer	0.54	0.07	-0.01	**0.52**	0.28
VideoMAE	**0.66**	0.07	0.21	0.50	0.36
Voice					
eGeMAPS v2	0.57	0.13	**0.44**	0.47	**0.40**
w2vBERT2	0.55	0.10	0.05	0.45	0.29
Text					
XLM RoBERTa	0.41	0.08	0.01	0.28	0.20

Table 3: Baseline selection across test datasets for NOXI-J trained models.

Feature Set	NOXI	NOXI-J	MPIGI	Additional	Combined
Video					
OpenFace 2.0	0.05	0.11	0.01	0.02	0.05
OpenPose	0.09	0.09	0.01	0.06	0.06
CLIP	0.04	0.26	0.07	0.04	0.10
DINO	0.11	0.20	-0.02	-0.04	0.06
SwinTransformer	0.16	0.22	-0.02	0.07	0.11
VideoMAE	**0.21**	0.11	**0.09**	**0.14**	**0.14**
Voice					
eGeMAPS v2	0.12	**0.30**	0.00	0.06	0.12
w2vBERT2	0.03	0.21	0.01	0.02	0.07
Text					
XLM RoBERTa	0.08	0.28	0.00	0.03	0.10

with eGeMAPS v2 close behind (0.12). Notably, eGeMAPS v2 excels in-language on NOXI-J (0.30) but transfers less to Additional/MPI than the NOXI-trained counterpart. Overall, audio features are the most consistent across domains (especially with NOXI training), while with NOXI-J training VideoMAE attains the highest Combined CCC.

Figure 1 summarizes CDD across languages and gender. In the case of CDD for languages, both the NOXI-trained (*left*) and NOXI-J–trained (*middle*) baselines show values clustered near zero (typically within ±0.05). The largest shifts arise for some vision backbones: for the NOXI-trained track we observe *VideoMAE* on JA/ZH ($\approx -0.11/-0.12$) and *DINO* on AR (≈ -0.10); for the NOXI-J–trained baselines, differences are more pronounced for *DINO* (IT ≈ -0.23, ZH $\approx +0.11$), *Swin* (EN/ES ≈ -0.15, ZH $\approx +0.12$), and *VideoMAE* (AR ≈ -0.12, DE $\approx +0.11$). In contrast, audio features (*eGeMAPS v2*, *w2vBERT2*) remain consistently close to zero across both training sets; text (*XLM-R*) shows only small shifts (e.g., EN ≈ -0.05, ZH $\approx +0.03$ in the NOXI-J–trained case). Concerning discrepancies with respect to gender, CDD$_G$ (*right*) is near zero for most features under both same-set conditions, with notable outliers for *SwinTransformer* on NOXI-J ($\approx +0.26$, higher predictions for males) and *CLIP* on NOXI (≈ -0.14). Overall, paralinguistic audio features are both accurate and comparatively fair, whereas high-capacity visual encoders exhibit modest, dataset and language-dependent sensitivity.

5 Challenge Solution Results

Table 4 reports the top-3 leaderboard for the multi-domain engagement estimation task. Systems are ranked by the *Combined* score, i.e., the mean CCC across the four test sets (NOXI base, NOXI (Additional), NOXI-J, MPIIGroupInteraction). The winning entry, HFUT–LMC [43], introduces *domain prompting via learnable adapters (DAPA)* to inject domain cues while preserving shared representations, and a *parallel cross-attention* module that aligns reactive (forward BiLSTM) and anticipatory (backward BiLSTM) states across interlocutors. This model achieves the best overall result with a Combined CCC of 0.699 and sets a new state of the art (SOTA) on NOXI test data 0.795 outperforming Li et al. [17]. The runner-up, USTC–IAT United [44], combines a BiLSTM + Transformer encoder with explicit target–partner fusion; an 8× overlapped sliding window pipeline and adaptive layer normalization further stabilize long-range regression. The third place, lasii, also exceeds the baseline with strong performance on NOXI and NOXI (Additional Languages).

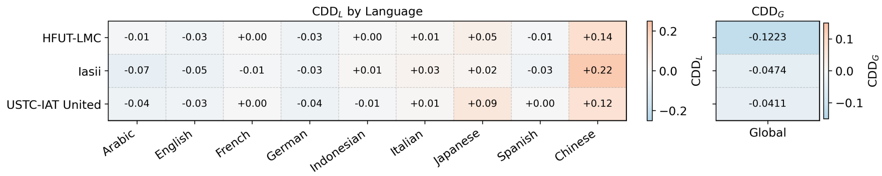

Figure 2: Language- and gender-wise fairness for all teams. Left: CDD$_L$ per language (lower magnitude is better). Right: global CDD$_G$ (male minus female), where values near zero indicate smaller gender disparity.

Table 4: Leaderboard for the multi-domain engagement estimation task (metric: CCC, higher is better). Combined is the mean across NOXI, NOXI (Additional Languages), NOXI-J, and MPIIGroupInteraction.

Team	NOXI	NOXI-J	MPIGI	Additional	Combined
Competition Teams					
HFUT-LMC	**0.795**	**0.578**	**0.668**	**0.755**	**0.699**
USTC-IAT United	0.788	0.530	0.664	0.732	0.678
lasii	0.789	0.512	0.544	0.732	0.644
Prior SOTA (MM 2024)					
DAT (Li et al [17])	0.760	—	0.490	0.670	—
Baseline					
Baseline (ours)	0.570	0.440	0.130	0.470	0.400

Fairness analysis (CDD). We evaluate conditional demographic disparity with respect to *language* (CDD$_L$) and *gender* (CDD$_G$). Figure 2 visualizes CDD$_L$ per language (blue = lower, red = higher predicted engagement relative to the bin-wise mean at the same ground truth). Most values cluster near zero, indicating modest language-related shifts overall; the largest deviations appear for *Chinese* and, to a lesser extent, *Japanese*. The right panel summarizes global CDD$_G$ per team: magnitudes are small, with HFUT–LMC showing the largest absolute shift (−0.1223) and USTC–IAT United and lasii closer to parity. Taken together with the CCC results, these patterns indicate that while leading methods improve accuracy, notable disparities persist, especially for Chinese across all systems and for gender in HFUT–LMC, highlighting the need for targeted analysis and mitigation in future work.

Other MultiMediate tasks. Beyond engagement estimation, we also hosted established tracks, Eye Contact, Bodily Behaviour Recognition, and Backchannel Detection from previous years challenges. On *Eye Contact*, USTC–IAT–United [42] set a new best test accuracy of 0.82, surpassing the MultiMediate '24 top result of 0.79. On *Bodily Behaviour Recognition*, USTC–IAT–United achieved 0.65, improving over the best result from MultiMediate '24 (0.63).

6 Conclusion

We extended engagement estimation to a cross-cultural, multi-domain setting by releasing NOXI-J (Japanese/Chinese) alongside NOXI, providing baselines trained on each, and evaluating on NOXI, NOXI (Additional Languages), NOXI-J, and MPIIGroupInteraction. Several teams surpassed the baselines across tasks (including engagement estimation). In aggregate, audio features exhibit the most consistent cross-language transfer, whereas visual encoders achieve strong in-domain accuracy with greater variability across domains. Fairness (CDD) magnitudes are generally small, with larger deviations for Chinese (and occasionally Japanese) and isolated gender effects. We release datasets, features, code, and leaderboards to support further work beyond MultiMediate '25.

Acknowledgments

The research conducted by Elisabeth André and Daksitha Withanage Don was partially funded by the German Research Foundation (DFG, SCHWAN project, Number 490909448), and Leibniz award of Elisabeth André under Grant AN 559/10-1. 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). M. Balazia was funded by the French National Research Agency under the UCA$^{\text{JEDI}}$ Investments into the Future, project number ANR-15-IDEX-01.

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] 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. doi:10.1145/3503161.3548363

[3] 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

[4] 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).

[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 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*. 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] 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

[9] 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

[10] 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

[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. doi: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). doi: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. doi:10.1007/978-3-030-78114-9_19

[16] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. doi:10.1073/pnas.1915768117

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

[19] 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

[20] 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

[21] 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

[22] 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.

[23] 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

[24] 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

[25] 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. doi:10.1145/3204493.3204549

[26] 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

[27] 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). doi:10.3389/frobt.2020.00092

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

[29] 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.

[30] 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

[31] 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.

[32] 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.

[33] 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

[34] 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

[35] 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.

[36] Dominik Schiller, Tobias Hallmen, Daksitha Withanage Don, Elisabeth André, and Tobias Baur. 2024. DISCOVER: A Data-driven Interactive System for Comprehensive Observation, Visualization, and ExploRation of Human Behaviour. *arXiv preprint arXiv:2407.13408* (2024).

[37] 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

[38] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law and Security Review* 41 (2021), 105567. doi:10.1016/j.clsr.2021.105567

[39] 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

[40] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue backchannel responses in English and Japanese. *Journal of Pragmatics* 32, 8 (July 2000), 1177–1207. doi:10.1016/S0378-2166(99)00109-5

[41] Daksitha Senel Withanage Don, Dominik Schiller, Tobias Hallmen, Silvan Mertes, Tobias Baur, Florian Lingenfelser, Mitho Müller, Lea Kaubisch, Prof. Dr. Corinna Reck, and Elisabeth André. 2024. Towards Automated Annotation of Infant-Caregiver Engagement Phases with Multimodal Foundation Models. In *Proc. of the ACM International Conference on Multimodal Interaction*. 428–438. doi:10.1145/3678957.3685704

[42] 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

[43] 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 *Proc. of the ACM International Conference on Multimedia*. doi:10.1145/3746027.3762079

[44] Yuefeng Zou, Hui Zhang, Jun Yu, Keda Lu, Linsi Zhu, Fengzhao Sun, Bo Wang, Kun Yao, Jianqing Sun, and Jiaen Liang. 2025. Heterogeneous Encoder Fusion with KAN Decoder for Group Engagement Modeling via 8× Sliding Pipelines. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery. doi:10.1145/3746027.3762078