BRAINIACS
BRAIN IMAGING AND COMPUTING SCIENCES
2025 Volume 6
Issue 3 Edoc F3DAD390E

Video Understanding of Complex Human Activities*

Michal Balazia†

Abstract

This report presents a comprehensive collection of novel methods and datasets advancing multimodal AI for behavioral analysis in social interactions, medical training, emotion recognition, and psychiatric phenotyping. Contributions include the MultiMediate'25 challenge for cross-cultural engagement estimation using expanded NOXI and NOXI-J corpora, a Go-ELAN YOLOv9 model for real-time surgical instrument detection in cataract videos, the CM3T adapter framework for efficient multimodal transfer learning on inhomogeneous datasets, the BLEMORE dataset with relative salience annotations for blended emotions, and MEPHESTO analyses revealing context-aware synchrony for therapeutic alliance, temporal variability for depression-schizophrenia classification, and trauma-modulated speech patterns in depression via MADRS/BDI-II assessments. These works of the INRIA-STARS team on video understanding of complex human activities bridge gaps in multimodal behavioral AI, supporting applications from assistive systems to personalized psychiatry.

Keyphrases

Video understanding, pattern recognition, human behavior analysis, social interactions, psychiatric phenotyping.

Contents

MultiMediate 2025: Cross-cultural multidomain engagement estimation

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 Challenge is an ACM Multimedia (ACMMM) Grand Challenge focused on multimodal group behavior analysis to move toward artificial mediation systems that can sense and support human interactions. Tasks include eye contact detection, bodily behavior recognition, backchannel detection, and engagement estimation, on the datasets MPIIGroupInteraction (Balazia et al. 2022) (see Figure 1) and NOXI (Cafaro et al. 2017). The most recent iteration, MultiMediate'25 (Müller et al. 2025), advances engagement estimation to more challenging, cross-cultural, and multidomain settings. Building on prior challenge editions (Müller et al. 2021; Müller et al. 2022; Müller et al. 2023; Müller et al. 2024), we expand beyond NOXI as the sole training source by introducing NOXI-J (Funk et al. 2024), a new multilingual corpus covering Japanese and Chinese interactions, enabling both training and evaluation in diverse linguistic contexts. MultiMediate'25 continues all previously defined tasks and creates another task: cross-cultural multidomain engagement estimation. MultiMediate'25 continues all previously defined tasks and creates another task: cross-cultural multidomain engagement estimation.

In Müller et al. (2025), we present new annotations, precomputed multimodal features (visual, vocal, and verbal), baseline evaluations, and an analysis of the best performing challenge solutions. Besides accuracy, we quantify fairness using conditional demographic disparity

*Report presented 2025-10-09, Guardians 2025, 4th BHAVI Guardians Conference.
†Correspondence to michal.balazia@inria.fr.

Table 1: Engagement estimation datasets used in the MultiMediate'25 challenge. Languages covered by each dataset are given in italics, with the respective number of interactions in parentheses.

Dataset	Training Data	Validation Data	Test Data
NOXI	*English (23), French (7), German (8)*	*English (3), French (4), German (3)*	*English (6), French (6), German (4), Arabic (2), Italian (2), Indonesian (4), Spanish (4)*
NOXI-J	*Japanese (21), Chinese (10)*	*Japanese (6), Chinese (4)*	*Japanese (6), Chinese (4)*
MPII Group		*German (6)*	*German (6)*

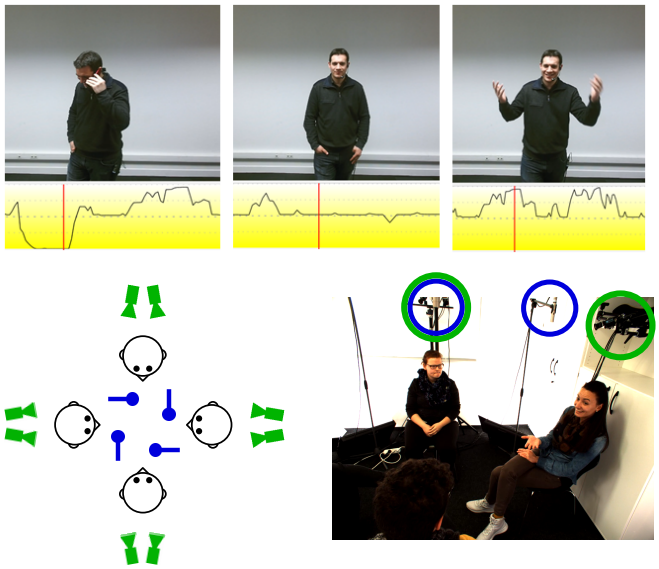

Figure 1: Top: Snapshots of scenes of a participant in the NOXI corpus being disengaged, neutral and highly engaged. Bottom: Setup of the MPIIGroupInteraction dataset.

(CDD) for gender and language. Our baselines confirm strong in-domain performance (e.g., paralinguistic GeMAPS (Eyben et al. 2015) and video-transformer features (Liu et al. 2022)) and reveal notable cross-domain drops, underscoring the challenge of cultural, linguistic, and interactional shifts. Fairness analyses indicate generally small CDD discrepancies clustered near zero (typically within ± 0.05) for both NOXI- and NOXI-J-trained baselines across languages and gender, with audio features showing the most consistency while high-capacity visual encoders exhibit modest, dataset- and language-dependent sensitivities. We observe the largest disparities for the proposed challenge solutions on the Chinese language part. All annotations, features, code, and leaderboards are made publicly available to foster sustained progress on robust and fair engagement estimation.

Participants are provided with the training datasets NOXI and NOXI-J. NOXI 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. 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 serve as an out-of-domain evaluation set. NOXI is extended by NOXI-J consisting of 66 dyadic interactions and over 16 hours of material using the same setup as original NOXI. NOXI-J features 48 interactions in Japanese with native Japanese speakers and 18 interactions in Chinese with Chinese native speakers. See Table 1 for the train-validation-test split.

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 labeled data for training and validation and undergo in-domain and out-of-domain evaluations on NoXI, NoXI-J, NoXI (Additional Languages), and MPIIGroupInteraction. We provide a multimodal set of precomputed features to participants. From the audio signal, we provide transcripts generated with the Whisper model. Additionally, we supply GeMAPS (Eyben et al. 2015) features along with wav2vec 2.0 embeddings (Barrault et al. 2023). From the video, we provide the backbone embeddings of Video Swin Transformer (Liu et al. 2022), DINOv2 (Oquab et al. 2024), CLIP (Radford, Kim, Hallacy, et al. 2021) and VideoMAEv2 (L. Wang et al. 2023) and the outputs of OpenFace 2.0 (Baltrusaitis et al. 2018) and OpenPose (Cao et al. 2019) to cover facial as well as body behaviors.

Identifying surgical instruments in pedagogical cataract surgery videos

Instructional cataract surgery videos are crucial for ophthalmologists and trainees to observe surgical details repeatedly. In Sinha et al. (2025), we present a deep learning model for real-time identification of surgical instruments in these videos, using a custom dataset scraped from open-access sources. Inspired by the architecture of YOLOv9 (C.-Y. Wang, Yeh, et al. 2025), the model employs a Programmable Gradient Information (PGI) mechanism and a novel Generally-Optimized Efficient Layer Aggregation Network (Go-ELAN) to address the information bottleneck problem, enhancing Minimum Average Precision (mAP) at higher Non-Maximum Suppression Intersection over Union (NMS IoU) scores.

Go-ELAN YOLOv9 Architecture (see Figure 2) contains an auxiliary block which works on the Programmable Gradient Information (PGI) concept by creating an auxiliary reverse branch for enabling reliable gradient calculation by avoiding potential semantic loss. The proposed Go-ELAN blocks use much larger filter sizes (512 instead of 128) to scan wider areas of each video frame at once. This captures the full context around tiny, fast-moving surgical tools (e.g., a needle glinting briefly under the microscope) without losing sight of their delicate tips, which is crucial in close-up cataract surgery footage where instruments constantly overlap and shift scales. Meanwhile, the Spatial Pyramid Pooling block frees the system from assuming all tools appear at fixed

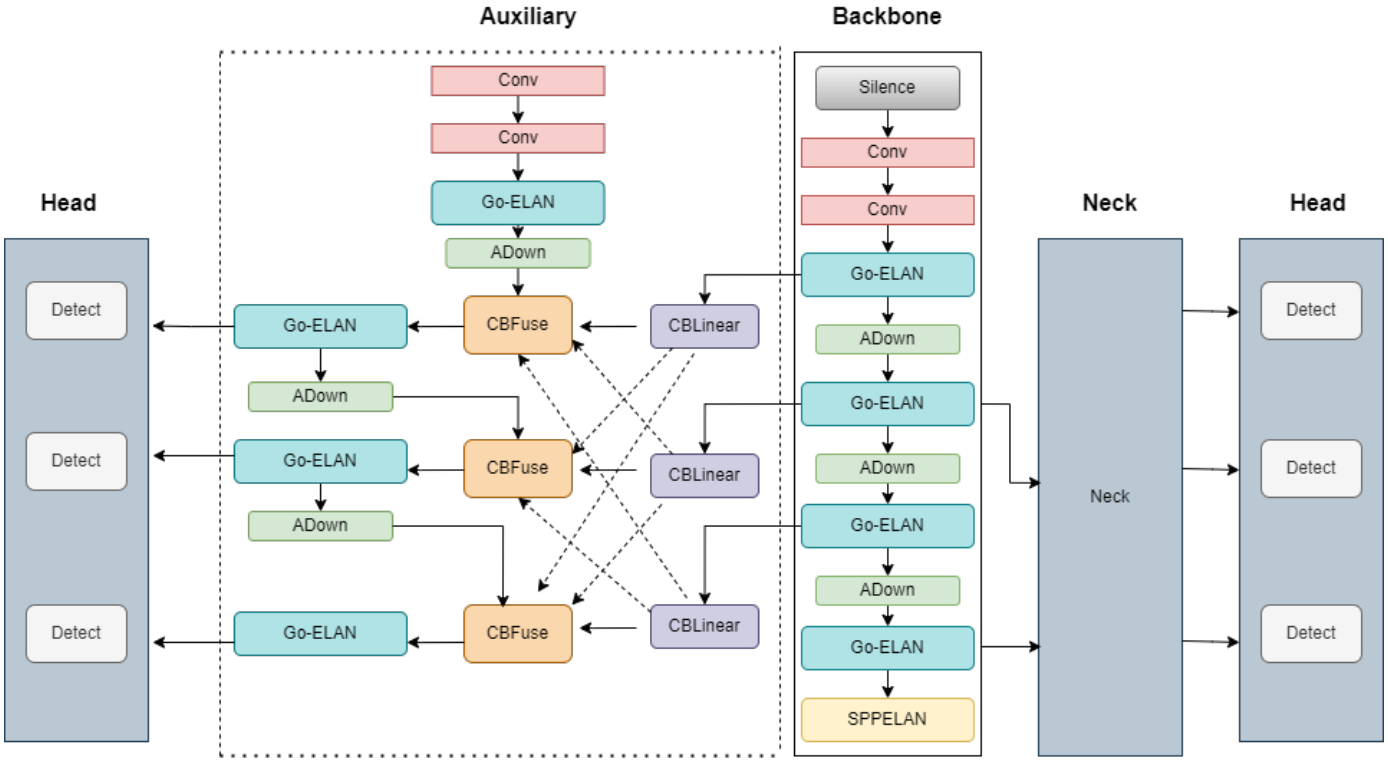

Figure 2: Architecture of Go-ELAN YOLOv9.

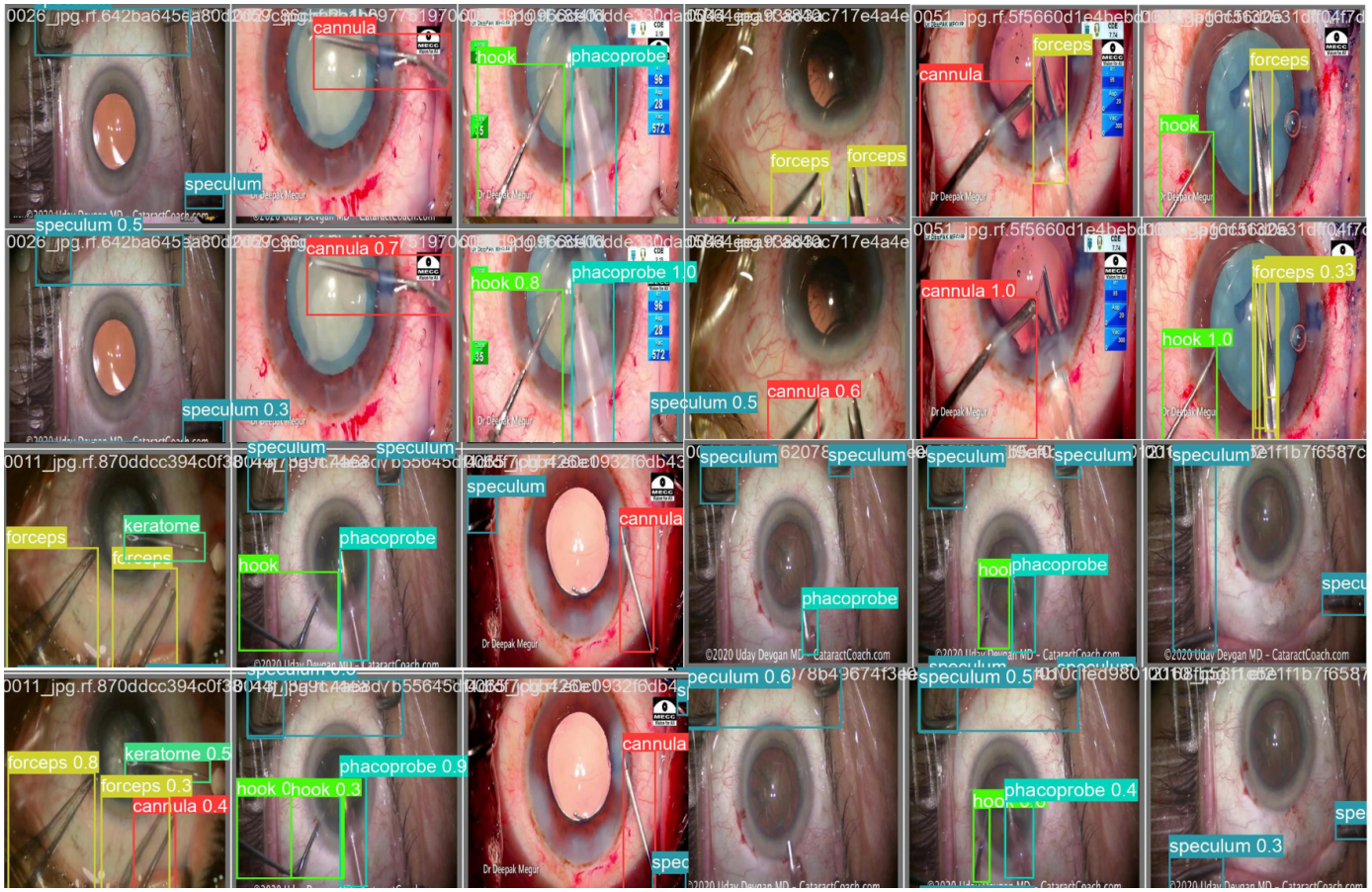

Figure 3: Qualitative examination of model performance. Rows 1 and 3 are labels while 2 and 4 are respective predictions.

Table 2: Numbers of recordings in the full BLEMORE dataset, split by single emotions and all pairs of emotion blends. For each blend, we also show the numbers of recordings split in all three (unequal 30% and 70%, equal 50% and 50%, unequal 70% and 30%) salience proportions.

Single Emotion	Blended Emotion	30/70	50/50	70/30
anger 226	anger+disgust	56	56	56
disgust 231	anger+fear	58	58	57
fear 236	anger+happiness	59	58	58
happiness 236	anger+sadness	56	57	57
sadness 233	disgust+fear	57	56	57
neutral 228	disgust+happiness	53	53	53
	disgust+sadness	54	54	54
	fear+happiness	53	54	54
	fear+sadness	53	53	53
	happiness+sadness	55	54	54

sizes, letting it handle both broad views of the surgical field and extreme zooms on fine instruments in a single pass. The neck then combines these multi-scale features before final head predictions.

Our Go-ELAN YOLOv9 model, evaluated against YOLOv5 (Jocher 2020), YOLOv7 (C.-Y. Wang, Bochkovskiy, et al. 2023), YOLOv8 (Jocher et al. 2023), vanilla YOLOv9 (C.-Y. Wang, Yeh, et al. 2025), Laptool (Namazi et al. 2022) and DETR (Carion et al. 2020), achieves a superior mAP of 73.74 at IoU 0.5 on a dataset of 615 images with 10 instrument classes, demonstrating the effectiveness of the proposed model. To illustrate the visual and qualitative superiority of our model, we have compared 12 ground-truth images of a cataract surgery with their respective model predictions in Figure 3.

CM3T framework for efficient multimodal learning for inhomogeneous interaction datasets

Challenges in cross-learning, such as domain shifts or negative transfer, involve inhomogeneous or even inadequate amount of training data and lack of resources for retraining large pretrained models. Inspired by transfer learning techniques in NLP, adapters and prefix tuning, we present a new model-agnostic plugin architecture for cross-learning, called CM3T (Agrawal et al. 2025), that adapts transformer-based models to new or missing information (see Figure 4). We introduce two adapter blocks: multi-head vision adapters for transfer learning and cross-attention adapters for multimodal learning. In comparison to training the full architecture, training only the adapters becomes substantially more efficient as the backbone and other plugins do not need to be fine tuned along with these additions.

CM3T has no specific requirements for training or pretraining and is a step towards bridging the gap between a general model and specific practical applications of video classification. Comparative and ablation studies on three datasets Epic-Kitchens-100 (Damen et al. 2020), MPIIGroupInteraction (Balazia et al. 2022) and UDIVA v0.5 (Palmero et al. 2021) show efficacy of this framework on different recording settings and tasks. With only 12.8% trainable parameters compared to the backbone to process video input and only 22.3% trainable parameters for two additional modalities, we achieve comparable and even better results than the state-of-the-art: 48.2% top-1 accuracy

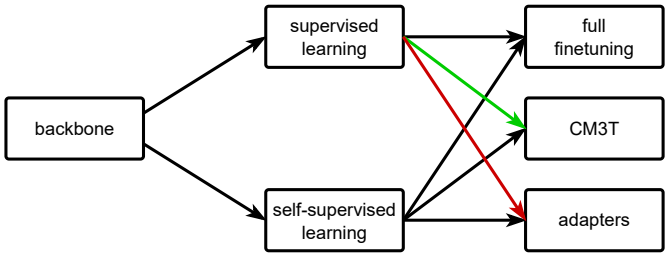

Figure 4: This diagram represents the main problem CM3T aims to solve. Backbones pretrained using self-supervised learning provide good general features, thus all methods of finetuning work well. In the case of supervised pretraining, adapters fail to perform well (in red) and CM3T is introduced to solve this (in green).

on Epic-Kitchens-100 compared to MM Mix's 49.6%, 0.901 mAP on MPIIGroupInteraction surpassing FAt Transformers with 0.899 and on UDIVA v0.5 with 0.69 MSE (vs. 0.72).

BLEMORE dataset of blended emotion expressions with relative salience annotations

Humans often experience not just a single basic emotion at a time, but rather a blend of several emotions with varying salience. Despite the importance of such blended emotions, most video-based emotion recognition approaches are designed to recognize single emotions only. The few approaches that have attempted to recognize blended emotions typically cannot assess the relative salience of the emotions within a blend. This limitation largely stems from the lack of datasets containing a substantial number of blended emotion samples annotated with relative salience. To address this shortcoming, we introduce BLEMORE (Lachmann et al. 2026), a novel dataset for multimodal (video, audio) BLended EMOtion REcognition that includes information on the relative salience of each emotion within a blend.

BLEMORE (see Figure 5) comprises over 3,000 clips from 58 actors, performing 6 basic emotions (anger, disgust, fear, happiness, sadness, and neutral) and 10 distinct blends consisting of all pairwise combinations of anger, disgust, fear, happiness, and sadness. All pairwise combinations were further conveyed with three different blend conditions (see Table 2):

- 50/50 = same amount of both emotions (e.g. 50/50 happiness-sadness, both happiness and sadness are expressed in equal proportions)
- 70/30 = the first emotion is more salient than the second emotion (e.g. 70/30 happiness-sadness conveys mainly happiness blended with a tinge of sadness)
- 30/70 = the second emotion is more salient than the first emotion (e.g. 30/70 happiness-sadness conveys mainly sadness blended with a tinge of happiness)

Labels are based directly on the actors' instructions for portraying specific single or blended emotions with defined salience ratios.

Using this dataset, we conduct extensive evaluations of state-of-the-art video classification approaches on two blended emotion prediction tasks: (1) predicting the presence of emotions in a given sample, and (2) predicting the relative salience of emotions in a blend. Our results

Figure 5: Examples of stills from the video recordings. The actor portrays a combination of anger and fear.

show that unimodal classifiers achieve up to 29% presence accuracy and 13% salience accuracy on the validation set, while multimodal methods yield clear improvements, with ImageBind (Girdhar et al. 2023) + WavLM (Chen et al. 2022) reaching 35% presence accuracy and HiCMAE (Sun et al. 2024) 18% salience accuracy. On the held-out test set, the best model VideoMAEv2 (L. Wang et al. 2023) + HuBERT (Hsu et al. 2021) achieves 33% presence accuracy and HiCMAE (Sun et al. 2024) 18% salience accuracy.

The BLEMORE dataset is also the basis of the BLEMORE competition where participants develop systems to predict the emotions present in each recording and the relative salience of each emotion. To support participation, we provide training data with labels, test data without labels, pre-extracted audio-visual feature embeddings, and baseline unimodal and multimodal classification results. The competition offers the first comprehensive platform for evaluating blended emotion recognition and aims to stimulate methodological innovation in multimodal affective computing.

MEPHESTO dataset for multimodal phenotyping of psychiatric disorders from social interaction

Identifying objective and reliable markers to tailor diagnosis and treatment of psychiatric patients remains a challenge, as conditions like major depression, bipolar disorder, or schizophrenia are qualified by complex behavior observations or subjective self-reports instead of easily measurable somatic features. Recent progress in computer vision, speech processing and machine learning has enabled detailed and objective characterization of human behavior in social interactions. However, the application of these technologies to personalized psychiatry is limited due to the lack of sufficiently large corpora that combine multimodal measurements with longitudinal assessments of patients covering more than a single disorder. Our multi-centre, multi-disorder longitudinal corpus creation effort MEPHESTO (König et al. 2022) is designed to develop and validate novel multimodal markers for psychiatric conditions. MEPHESTO consists of multimodal audio, video, and physiological recordings as well as clinical assessments of psychiatric patients covering a six-week main study period as well as several follow-up recordings spread across twelve months.

Diagnoses include schizophrenia, depression and bipolar disorder. The dataset is confidential and does not include control subjects. Each patient contributed with 1–8 videos, roughly 5.5 videos on average. In addition to video, the recordings include patients' and clinicians' biosignals electrodermal activity, blood volume pulse, inter-beat interval, heart rate, temperature, and hand acceleration. Videos are recorded by Azure Kinect and biosignals by Empatica. People do not wear face masks while being recorded. However, to minimize the transmission of COVID-19, there is a large transparent plexi-glass. Figure 6 shows a screenshot from a mock recording. We have made three contributions regarding therapeutic alliance, recognizing depression and schizophrenia, and detecting childhood trauma from speech. These contributions are explained in detail in the subsections below.

Contextualized synchrony for therapeutic alliance

Non-verbal behavioral synchrony has been widely studied as an indicator of relational dynamics in clinical interactions (Jennissen et al. 2024) and has been shown to exhibit weak to moderate associations with therapeutic alliance (TA). However, most existing synchrony measures are computed in a content-agnostic manner, implicitly assuming that synchrony occurring at different moments of an interaction contributes equally to the development of the therapeutic relationship (Jennissen et al. 2024). This work is motivated by the hypothesis that the relational meaning of synchrony is context-dependent, and that linguistic content may play a critical role in determining when non-verbal coordination is most relevant to therapeutic alliance. In our setting, TA is assessed at the end of each session via a seven-item patient questionnaire (König et al. 2022) capturing liking, perceived helpfulness, feeling understood and supported, and ease of sharing personal information, with the global TA score obtained by averaging item responses. By integrating semantic information derived from spoken language with non-verbal synchrony measures, this study aims to move beyond global, uniform synchrony metrics toward a more fine-grained, context-sensitive understanding of therapist–patient interaction dynamics. Non-verbal synchrony was computed at the window level using Motion Energy Analysis (MEA by Ramseyer and Tschacher (2011)). Figure 7 displays an example of patient–therapist MEA time series and a cross-correlation analysis applied to the continuous motion energy time series of patient and therapist.

For evaluations, we take a subset of the MEPHESTO dataset containing 106 pairs of patient-clinician videos. We evaluate all models by predicting session-level TA scores and using Pearson's correlation coefficient r between predicted and observed TA as the primary outcome measure, computed in a session-level cross-validation setting. We first replicated a stable baseline association between global MEA synchrony and patient-reported TA, with a content-agnostic aggregation over all windows yielding a correlation of approximately $r \approx 0.22$. Building on this foundation, transcript data were processed into semantic embeddings (dense vector representations of linguistic units) and temporally aligned with synchrony windows, enabling a multimodal representation in which textual context modulates how window-level synchrony is ag-

Figure 6: Screenshot of a mock recording with two videos and biosignals. The person on the left represents a clinician and the person on the right a patient. To protect the identity of patients, clinicians enact both roles in this mock recording.

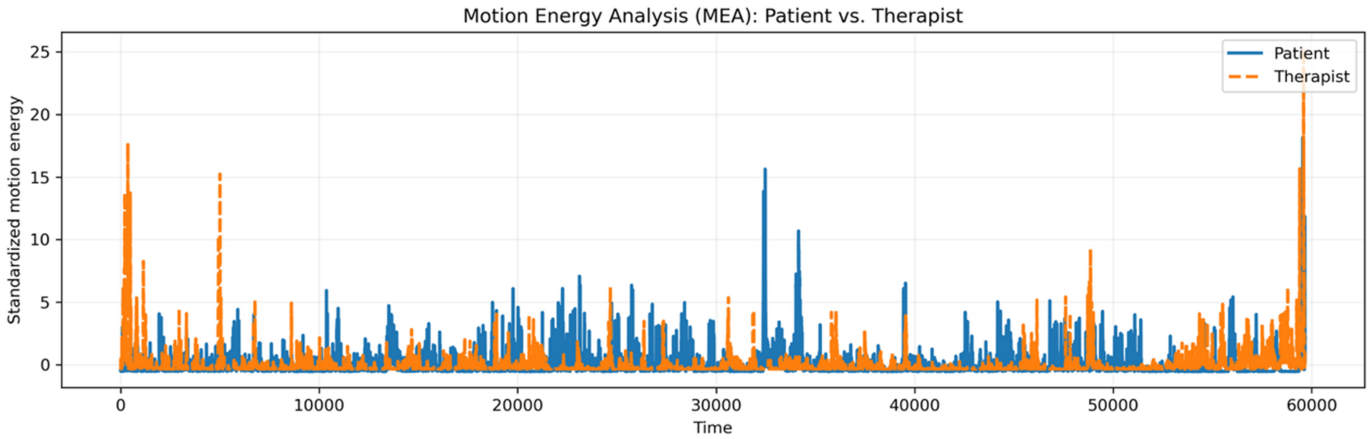

Figure 7: Example of patient–therapist Motion Energy Analysis (MEA) time series over a single therapy session.

gregated over time. In the current implementation, not all MEA windows have a corresponding text segment, so windows without aligned transcripts are ignored when applying text-informed weighting. Evaluating a uniform (all-ones) aggregation under this constraint leads to a reduced MEA-TA association of $r \approx 0.13$, compared to the $r \approx 0.22$ obtained when all MEA windows are used. Within this constrained evaluation setting, however, our text-informed weighting scheme increases the correlation to $r \approx 0.18$, suggesting that linguistic information helps to highlight synchrony segments that are more informative about alliance. While the overall performance of this preliminary implementation does not yet surpass the full-window MEA baseline, the results support the view that synchrony is not uniformly informative throughout an interaction and highlight the potential of window-level, context-aware multimodal modeling combined with improved textual coverage for capturing subtle relational processes in therapeutic settings.

Psychiatric diagnosis classification through temporal behavioral analysis

This part of the project focuses on automated psychiatric diagnosis through multimodal behavioral analysis of clinical interview videos,

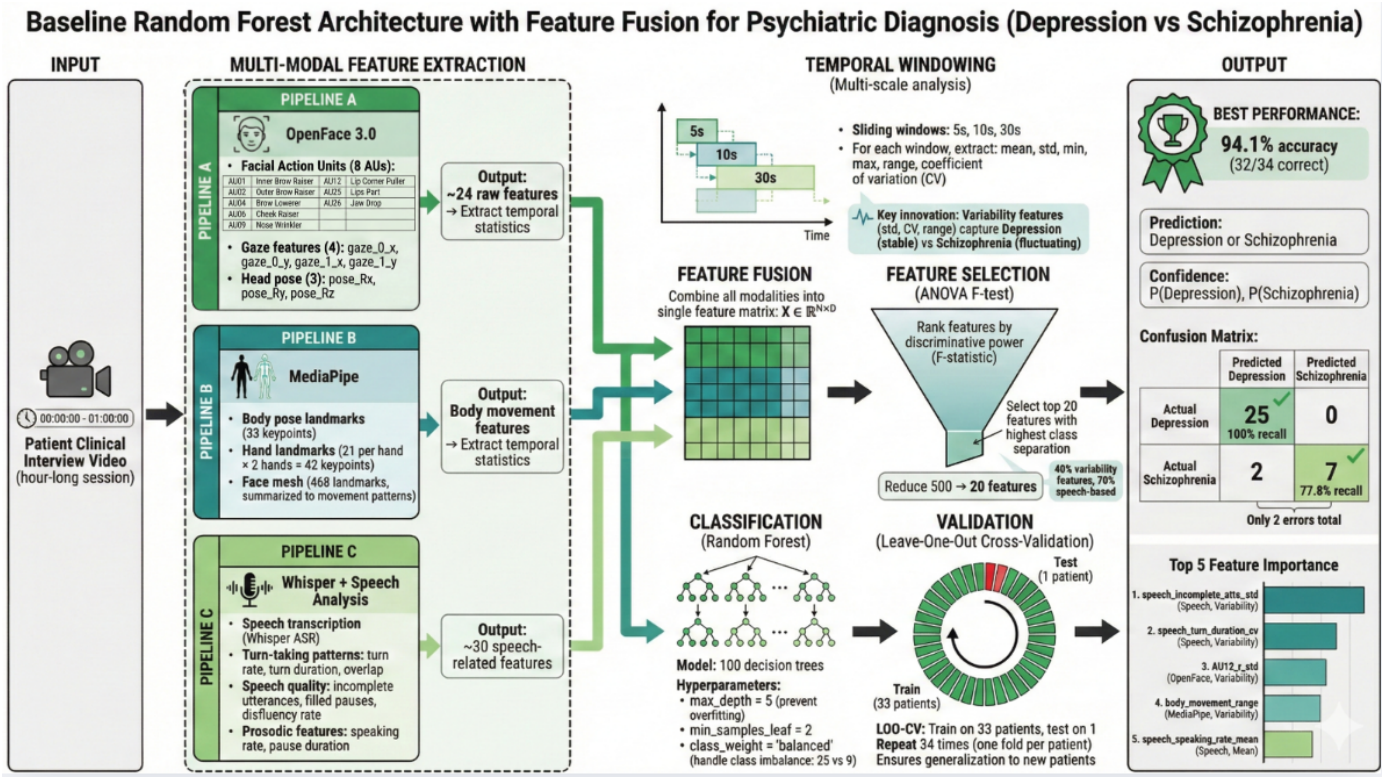

Figure 8: This architecture diagram illustrates a multimodal machine learning pipeline for binary psychiatric diagnosis (depression/schizophrenia) from clinical interview videos. The system combines three parallel feature extraction pipelines: OpenFace 3.0 for facial action units and gaze, MediaPipe for body pose and hand movements, and Whisper for speech transcription and linguistic analysis. Features are extracted across multi-scale temporal windows with statistical aggregations to capture temporal variability patterns. After feature fusion into a unified matrix, ANOVA F-test ranks features by discriminative power, selects the top 20, and makes predictions using a random forest classifier.

with the objective of distinguishing between depression and schizophrenia. We utilize a portion of the MEPHESTO dataset of 34 patients: 25 with depression and 9 with schizophrenia. The dataset includes manual behavioral annotations provided by expert clinical annotators who labeled over 3000 video segments with observable behaviors. The implemented system (see Figure 8) follows a 7-stage pipeline: (1) input data acquisition from MEPHESTO with pre-annotated transcriptions, (2) low-level extraction using OpenFace 3.0 (Hu et al. 2025) (8 action units + gaze + head pose + 8 emotions), MediaPipe holistic (Lugaresi et al. 2019) (33 pose, 42 hand, 468 face landmarks), and Whisper (Radford, Kim, Xu, et al. 2023) for speech (1,842 features/frame), (3) temporal alignment with frame-level synchronization (±1 frame precision, 33ms), (4) multi-scale windowing (5s, 10s, 30s windows, 50% overlap) extracting 188 features across 24,588 windows, (5) temporal variability aggregation computing 6 statistics (mean, standard deviation, coefficient of variation, minimum, maximum, range) per feature, (6) feature selection via ANOVA F-test selecting top 20 features (70% speech-based, 30% visual), and (7) classification with random forest using leave-one-out cross-validation across 13 tested methods. Random forest achieves 94.1% accuracy with only two schizophrenia patients misclassified. Top discriminative feature is the standard deviation of patient's incomplete utterances. During our experiments, we found that fluctuations in behavioral patterns are the critical discriminative marker, that speech features dominate (70%) in the top-20 features, that feature fusion outperforms modality separation, and that traditional machine learning beats deep learning on small datasets (Gimeno-Gómez et al. 2024). In the future, we are going to focus on temporal trauma detection in the long, untrimmed clinical interviews.

Childhood trauma affects speech and language measures in patients with major depressive disorder

Speech analysis has shown significant promise as a potential biomarker for depression (Koops et al. 2023). Childhood trauma may impact speech and language patterns in individuals with depression (Ettore et al. 2025). This study aims to explore the relationship between vocal characteristics and depressive symptoms, while also assessing how childhood trauma may shape these patterns. Twenty-seven adult MEPHESTO participants with a major depressive episode were included. The severity of depression was assessed using the Montgomery & Åsberg Depression Rating Scale (MADRS) (Montgomery and Åsberg 1979) and the Beck Depression Inventory II (Beck et al. 1996). Childhood trauma was measured using a childhood trauma questionnaire. Speech recordings from the MADRS semi-structured interview and a free clinical interview were analyzed using speaker diarization, automatic speech recognition, and feature extraction by Whisper (Radford, Kim, Xu, et al. 2023). We used Spearman's rank correlations ($p < 0.05$, no multiple testing correction) to identify acoustic features such as the second formant frequency SD ($\rho = -0.58, p = 0.002$) and the first Mel-Frequency Cepstral Coefficient ($\rho = 0.56, p = 0.002$) significantly associated with depression severity on Beck Depression Inventory II. Correlation analysis revealed that greater depression severity was linked to shorter, less diverse speech, characterized by fewer words, fewer

semantic clusters, and reduced articulatory effort. In contrast, childhood trauma was positively associated with distinct speech characteristics. Higher trauma load was associated with richer, longer, and more syntactically complex speech. Additionally, utterances were shorter, with more frequent shifts between semantic clusters, reflecting a more fragmented speech pattern influenced by traumatic load. Our study highlights the influence of childhood trauma on vocal and linguistic characteristics of patients with depression. Automated language analysis offers the possibility to identify biomarkers of traumatic load in patients with depression. This could improve diagnostic accuracy, guide therapeutic management and monitor clinical progress.

Conclusion

This report presents a comprehensive collection of INRIA-STARS projects of 2025 advancing multimodal AI for behavioral analysis across social interactions, surgical training, emotion recognition, and psychiatric phenotyping through innovative methods and datasets like MultiMediate25, Go-ELAN YOLOv9, CM3T, BLEMORE, and MEPHESTO analyses. Key contributions include cross-cultural engagement estimation on expanded NOXI/NOXI-J corpora, real-time surgical instrument detection , efficient CM3T adapters matching SOTA with minimal parameters, BLEMORE's salience-annotated blended emotions, and MEPHESTO's context-aware synchrony, temporal variability for diagnosis, and trauma-modulated speech biomarkers. These efforts bridge gaps in video understanding of complex human activities, demonstrating domain-informed features outperform neural models on limited clinical data. Public resources such as annotations, features, and code facilitate community progress in culturally robust engagement models and blended emotion prediction. Future extensions to multi-site validation and longitudinal phenotyping promise transformative impacts on healthcare AI.

Citation

Brainiacs Journal 2025 Volume 6 Issue 3 Edoc F3DAD390E.
Title: "Video Understanding of Complex Human Activities".
Authors: Michal Balazia.
Dates: created 2025-10-02, presented 2025-10-09, updated 2025-12-13, published 2025-12-13, revised 2026-05-27

Contact: michal.balazia@inria.fr
NPDS: LINKS/Brainiacs/Balazia2025VUCHA
DOI: 10.48085/F3DAD390E

Acknowledgments

This work has been a joint effort of multiple members of the STARS team at INRIA Universite Cote d'Azur as well as many of our associates. We gratefully acknowledge the contributions of Tanay Agrawal, Jan Alexandersson, Elisabeth Andre, Michel Benoit, Francois Bremond, Andreas Bulling, Daksitha Withanage Don, Eric Ettore, Marius Funk, Mohammed Guermal, Rene Hurlemann, Alexandra Konig, Tim Lachmann, Petri Laukka, Hali Lindsay, Philipp Muller, Alexandra Israelsson, Shogo Okada, Danilo Postin, Huajian Qiu, Philippe Robert, Miriana Russo, Teimuraz Saghinadze, Aowen Shi, Sanya Sinha, Christina Tornberg, and Johannes Troger. The research was funded by the French National Research Agency under the projects ANR-15-IDEX-01 and ANR-23-IACL-0001.

Affiliations

INRIA Université Côte d'Azur, Sophia Antipolis, France.

References

[1] T. Agrawal, M. Guermal, M. Balazia, and F. Bremond. "CM3T: Framework for Efficient Multimodal Learning for Inhomogeneous Interaction Datasets." In: *IEEE Xplore*. Preprint. Final paper accepted at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, February, 2025. 10 pages. IEEE/CVF. Tucson, United States, Feb. 2025. URL: https://hal.science/hal-04880258 (cited p. 4).

[2] M. Balazia, P. Müller, Á. L. Tánczos, A. v. Liechtenstein, and F. Brémond. "Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation." In: *Proc. of the ACM International Conference on Multimedia*. 2022, pp. 70–79. DOI: 10.1145/3503161.3548363 (cited pp. 1, 4).

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. "Openface 2.0: Facial behavior analysis toolkit." In: *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE. 2018, pp. 59–66. DOI: 10.1109/FG.2018.00019 (cited p. 2).

[4] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, et al. "Seamless: Multilingual Expressive and Streaming Speech Translation." *arXiv preprint arXiv:2312.05187* (2023) (cited p. 2).

[5] A. T. Beck, R. A. Steer, and G. Brown. "Beck Depression Inventory–II." *APA PsycTests* (1996). DOI: 10.1037/t00742-000 (cited p. 7).

[6] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. F. Valstar. "The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions." In: *Proc. of the International Conference on Multimodal Interaction*. 2017. DOI: 10.1145/3136755.3136780 (cited p. 1).

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cited p. 2).

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. "End-to-End Object Detection with Transformers." In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 213–229. ISBN: 978-3-030-58452-8 (cited p. 4).

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, et al. "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing." *IEEE Journal of Selected Topics in Signal Processing* 16.6 (Oct. 2022), pp. 1505–1518. ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3188113. URL: http://dx.doi.org/10.1109/JSTSP.2022.3188113 (cited p. 5).

[10] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020) (cited p. 4).

[11] E. Ettore, H. Lindsay, J. Tröger, M. Balazia, et al. "Childhood trauma affects speech and language measures in patients with major depressive disorder during clinical interviews." *Journal of Affective Disorders* 388 (Nov. 2025), p. 119769. ISSN: 0165-0327. DOI: 10.1016/j.jad.2025.119769 (cited p. 7).

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." *IEEE Transactions on Affective Computing* 7.2 (2015), pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417 (cited p. 2).

[13] M. Funk, S. Okada, and E. André. "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*. 2024, pp. 224–233. ISBN: 9798400704628. DOI: 10.1145/3678957.3685757 (cited p. 1).

[14] D. Gimeno-Gómez, A.-M. Bucur, A. Cosma, C.-D. Martínez-Hinarejos, and P. Rosso. "Reading Between the Frames: Multi-modal Depression Detection in Videos from Non-verbal Cues." In: *Advances in Information Retrieval*. Springer Nature Switzerland, 2024, pp. 191–209. ISBN: 9783031560279. DOI: 10.1007/978-3-031-56027-9_12 (cited p. 7).

[15] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. "ImageBind: One Embedding Space To Bind Them All." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. URL: https://facebookresearch.github.io/ImageBind (cited p. 5).

[16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021. URL: https://arxiv.org/abs/2106.07447 (cited p. 5).

[17] J. Hu, L. Mathur, P. P. Liang, and L.-P. Morency. "OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis." In: *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*. 2025, pp. 1–11. DOI: 10.1109/FG61629.2025.11099277 (cited p. 7).

[18] S. Jennissen, J. Huber, B. Ditzen, and U. Dinger. "Association between nonverbal synchrony, alliance, and outcome in psychotherapy: Systematic review and meta-analysis." *Psychotherapy Research* 35.7 (Nov. 2024), pp. 1213–1228. ISSN: 1468-4381. DOI: 10.1080/10503307.2024.2423662 (cited p. 5).

[19] G. Jocher. *Ultralytics YOLOv5*. Version 7.0. 2020. DOI: 10.5281/zenodo.3908559. URL: https://github.com/ultralytics/yolov5 (cited p. 4).

[20] G. Jocher, A. Chaurasia, and J. Qiu. *Ultralytics YOLOv8*. Version 8.0.0. 2023. URL: https://github.com/ultralytics/ultralytics (cited p. 4).

[21] A. König, P. Müller, J. Tröger, H. Lindsay, et al. "Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study." *Personalized Medicine in Psychiatry* 33-34 (July 2022), p. 100094. DOI: 10.1016/j.pmip.2022.100094. URL: https://hal.inria.fr/hal-03724844 (cited p. 5).

[22] S. Koops, S. G. Brederoo, J. N. de Boer, F. G. Nadema, A. E. Voppel, and I. E. Sommer. "Speech as a Biomarker for Depression." *CNS & Neurological Disorders - Drug Targets* 22.2 (Feb. 2023), pp. 152–160. ISSN: 1871-5273. DOI: 10.2174/1871527320666211213125847 (cited p. 7).

[23] T. Lachmann, P. Müller, T. Saghinadze, M. Balazia, A. Israelsson, and P. Laukka. "Not all Blends are Equal: The BLEMORE Dataset of Blended Emotion Expressions with Relative Salience Annotations." In: *Proceedings of the 20th IEEE International Conference on Automatic Face and Gesture Recognition*. 2026 (cited p. 4).

[24] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. "Video Swin Transformer." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 3192–3201. DOI: 10.1109/CVPR52688.2022.00320 (cited p. 2).

[25] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, et al. "MediaPipe: A Framework for Perceiving and Processing Reality." In: *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*. 2019. URL: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf (cited p. 7).

[26] S. A. Montgomery and M. Åsberg. "A New Depression Scale Designed to be Sensitive to Change." *British Journal of Psychiatry* 134.4 (1979), pp. 382–389. DOI: 10.1192/bjp.134.4.382 (cited p. 7).

[27] P. Müller, M. Balazia, T. Baur, M. Dietz, et al. "MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions." In: *Proceedings of the 31st ACM International Conference on Multimedia*. MM '23. Ottawa ON, Canada: Association for Computing Machinery, 2023, pp. 9640–9645. ISBN: 9798400701085. DOI: 10.1145/3581783.3613851. URL: https://doi.org/10.1145/3581783.3613851 (cited p. 1).

[28] P. Müller, M. Balazia, T. Baur, M. Dietz, et al. "MultiMediate'24: Multi-Domain Engagement Estimation." In: *Proceedings of the 32nd ACM International Conference on Multimedia*. MM '24. Melbourne VIC, Australia: Association for Computing Machinery, 2024, pp. 11377–11382. ISBN: 9798400706868. DOI: 10.1145/3664647.3689004. URL: https://doi.org/10.1145/3664647.3689004 (cited p. 1).

[29] P. Müller, M. Dietz, D. Schiller, D. Thomas, H. Lindsay, P. Gebhard, E. André, and A. Bulling. "MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions." In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22. Lisboa, Portugal: Association for Computing Machinery, 2022, pp. 7109–7114. ISBN: 9781450392037. DOI: 10.1145/3503161.3551589. URL: https://doi.org/10.1145/3503161.3551589 (cited p. 1).

[30] P. Müller, M. Dietz, D. Schiller, D. Thomas, G. Zhang, P. Gebhard, E. André, and A. Bulling. "MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation." In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. Virtual Event, China: Association for Computing Machinery, 2021, pp. 4878–4882. ISBN: 9781450386517. DOI: 10.1145/3474085.3479219. URL: https://doi.org/10.1145/3474085.3479219 (cited p. 1).

[31] B. Namazi, G. Sankaranarayanan, and V. Devarajan. "A contextual detector of surgical tools in laparoscopic videos using deep learning." *Surgical Endoscopy* 36.1 (2022), pp. 679–688. ISSN: 1432-2218. DOI: 10.1007/s00464-021-08336-x. URL: https://doi.org/10.1007/s00464-021-08336-x (cited p. 4).

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. URL: https://arxiv.org/abs/2304.07193 (cited p. 2).

[33] C. Palmero, J. Selva, S. Smeureanu, J. Junior, et al. "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1–12 (cited p. 4).

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, et al. "Learning Transferable Visual Models From Natural Language Supervision." In: *International Conference on Machine Learning*. 2021. URL: https://api.semanticscholar.org/CorpusID:231591445 (cited p. 2).

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. "Robust speech recognition via large-scale weak supervision." In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023 (cited p. 7).

[36] F. T. Ramseyer and W. Tschacher. "Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome." *Journal of consulting and clinical psychology* 79 3 (2011), pp. 284–95. URL: https://api.semanticscholar.org/CorpusID:32001201 (cited p. 5).

[37] S. Sinha, M. Balazia, and F. Bremond. "Identifying Surgical Instruments in Pedagogical Cataract Surgery Videos through an Optimized Aggregation Network." In: *Image Processing Applications and Systems*. IEEE. Lyon, France, Jan. 2025. URL: https://inria.hal.science/hal-04864972 (cited p. 2).

[38] L. Sun, Z. Lian, B. Liu, and J. Tao. "Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition." *Information Fusion* 108 (2024), p. 102382 (cited p. 5).

[39] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors." In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 7464–7475. DOI: 10.1109/CVPR52729.2023.00721 (cited p. 4).

[40] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information." In: *Computer Vision – ECCV 2024*. Ed. by A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol. Cham: Springer Nature Switzerland, 2025, pp. 1–21. ISBN: 978-3-031-72751-1 (cited pp. 2, 4).

[41] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. "VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking." In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14549–14560. DOI: 10.1109/CVPR52729.2023.01398 (cited pp. 2, 5).

[42] D. Withanage, S. Daksitha, M. Funk, M. Balazia, et al. "MultiMediate '25: Cross-cultural Multi-domain Engagement Estimation." In: *Proceedings of the 33rd ACM International Conference on Multimedia*. MM '25. Dublin, Ireland: Association for Computing Machinery, 2025, pp. 14150–14155. ISBN: 9798400720352. DOI: 10.1145/3746027.3762076. URL: https://doi.org/10.1145/3746027.3762076 (cited p. 1).

