

A Multimodal Cross-Attention Framework for Robot Error Detection from Bystander Reactions

GERING, Paul <<http://orcid.org/0009-0009-5582-2718>>, TIAN, Ruyi <<http://orcid.org/0009-0009-5534-4832>>, CHARDIWALL, Samiulhaq <<http://orcid.org/0009-0003-5741-6154>>, WANG, Yuchen <<http://orcid.org/0000-0002-6352-8633>>, DI NUOVO, Alessandro <<http://orcid.org/0000-0003-2677-2650>>, CHEN, Chaona <<http://orcid.org/0000-0001-7289-1242>>, CHEN, Zhixiang <<http://orcid.org/0000-0002-5636-6082>> and MOORE, Roger K <<http://orcid.org/0000-0003-0065-3311>>

Available from Sheffield Hallam University Research Archive (SHURA) at:

<https://shura.shu.ac.uk/38058/>

This document is the Published Version [VoR]

Citation:

GERING, Paul, TIAN, Ruyi, CHARDIWALL, Samiulhaq, WANG, Yuchen, DI NUOVO, Alessandro, CHEN, Chaona, CHEN, Zhixiang and MOORE, Roger K (2026). A Multimodal Cross-Attention Framework for Robot Error Detection from Bystander Reactions. In: ROSSI, Silvia, ESPOSITO, Anna, TORTORA, Genoveffa, CONATI, Cristina, CUTUGNO, Francesco and SIEGERT, Ingo, (eds.) ICMI '26: Proceedings of the 28th International Conference On Multimodal Interaction. ACM, 1386-1390. [Book Section]

Copyright and re-use policy

See <http://shura.shu.ac.uk/information.html>

A Multimodal Cross-Attention Framework for Robot Error Detection from Bystander Reactions

Paul Gering
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
prgering1@sheffield.ac.uk

Ruyi Tian
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
rtian14@sheffield.ac.uk

Samiulhaq Chardiwall
School of Computing
Sheffield Hallam University
Sheffield, South Yorkshire, United Kingdom
samiulhaq.chardiwall@student.shu.ac.uk

Yuchen Wang
University of Sheffield
Sheffield, South Yorkshire, United Kingdom
ywang1016@sheffield.ac.uk

Alessandro Di Nuovo
School of Computing
Sheffield Hallam University
Sheffield, United Kingdom
a.dinuovo@shu.ac.uk

Chaona Chen
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
chaona.chen@sheffield.ac.uk

Zhixiang Chen
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
zhixiang.chen@sheffield.ac.uk

Roger K. Moore
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
r.k.moore@sheffield.ac.uk

Abstract

To detect human-robot interaction errors, existing approaches have primarily focused on behavioural responses from individuals directly involved in the interaction, whereas reactions from external observers have received comparatively little attention. Similarly, these approaches have been limited to interactions in controlled environments with little variability in lighting, camera angles, and background noise. Addressing these gaps, the ERR@HRI 3.0 challenge focuses on bystander error detection in unconstrained, real-world environments. Our primary submission is a time-aware multimodal fusion framework that combines facial and acoustic representations via bidirectional cross-attention. Two versions of this framework, with different encoding approaches, are evaluated against a bidirectional LSTM (BiLSTM) classifier with simple concatenation. Our results demonstrate the effectiveness of this cross-fusion approach for capturing errors and generalising to unseen real-world data. Our cross-fusion and BiLSTM models achieved the top-three positions on Track 1, outperforming the challenge baseline and other competing submissions. Feature importance analysis further revealed that smile-related action units and their temporal dynamics serve as primary discriminative cues for error detection. Our code is available at <https://github.com/chardiwall/SHEF-HRI>.

CCS Concepts

• Human-centered computing → HCI design and evaluation methods; • Computing methodologies → Neural networks.



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICMI '26, Napoli, Italy*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
<https://doi.org/10.1145/3776574.3832482>

Keywords

human-robot interaction, bystander error detection, cross-attention, multimodal fusion

ACM Reference Format:

Paul Gering, Ruyi Tian, Samiulhaq Chardiwall, Yuchen Wang, Alessandro Di Nuovo, Chaona Chen, Zhixiang Chen, and Roger K. Moore. 2026. A Multimodal Cross-Attention Framework for Robot Error Detection from Bystander Reactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3776574.3832482>

1 Introduction

Robotic systems often make errors, such as generating inappropriate or delayed responses, which can negatively impact the quality of human-robot interaction (HRI) [2]. To enable robots to identify and recover from these errors, recent research has explored methods for automatic error detection, with a focus on leveraging human behavioural responses as rupture indicators [7, 12, 13]. The ERR@HRI challenge series has been instrumental in driving this research, providing a standardised framework for developing and evaluating machine-learning models in this domain [2, 11].

While the fundamental objective of the ERR@HRI challenge has remained consistent, its scope has evolved. The first two iterations evaluated dyadic HRI in controlled laboratory settings, training models on multimodal features, including facial action units, prosody and body posture, to predict interaction ruptures [2, 11]. In contrast, the third iteration (ERR@HRI 3.0) shifts the focus towards a third-party perspective, utilising recorded bystander reactions to predict errors (Track 1) and the anticipated outcome of the interaction (Track 2) [8]. These bystander recordings are captured in uncontrolled environments, with variable lighting, camera

angles, and environmental noise, making multimodal error detection considerably more difficult.

Given the unconstrained nature of the challenge data, proposed solutions need to prioritise generalisability. Previous challenge submissions, which relied on simple concatenation to fuse multimodal features, exhibited sharp performance drops when moving from validation to the test sets [7, 9]. We, therefore, argue that simple feature concatenation is insufficient for capturing complex cross-modal dynamics and that more sophisticated fusion steps are required to handle the environmental variability in the ERR@HRI 3.0 data.

In this paper, we propose a time-aware multimodal fusion approach capable of modelling the temporal dynamics within modalities and the interactions between them. Our approach adapts the cross-modal transformer by Venkatraman et al. [4], which uses bidirectional cross-attention to allow each modality to query the other for temporally aligned representations. We train these cross-attention models on facial expression and acoustic features extracted from the challenge dataset. Model performance is evaluated against a bidirectional LSTM (BiLSTM) with simple concatenation and the official challenge baseline (BadNetCNN) for Bystander Affect Detection (Track 1). We focus on Track 1 as it addresses the immediate connection between interaction errors and bystander reactions while also presenting key challenges such as severe class imbalance. Finally, we conduct a discriminative feature analysis to identify which multimodal cues contribute most to error detection.

2 ERR@HRI Challenge

2.1 Datasets

The ERR@HRI 3.0 challenge [8] provides two datasets, both released as raw, non-anonymised webcam recordings (.mp4) collected via Prolific, which preserves real-world variability in lighting, camera angle, and participant positioning. We focus on the BAD (Bystander Affect Detection) dataset [1] (Track 1), which comprises 45 participants' spontaneous facial reactions to 46 stimuli, each labelled as Failure or Control, with a strong imbalance towards the Failure class (86.9% of clips). We used the official subject-independent splits: 36 participants (1,319 videos) for training and validation, and 9 participants (326 videos) for test, with the Failure-class proportion (~87%) preserved across both.

2.2 Evaluation Protocol

Track 1 is a binary classification task over a held-out, subject-independent test split, with each (participant, video) pair treated as an independent observation. Predictions are made at the window level: each recording is segmented into fixed-length sliding windows, subject to a 2-second window cap. Window-level predictions are then aggregated to a video-level decision by majority vote, with ties resolved in favour of Failure. The task is ranked by video-level Macro F_1 -score with balanced accuracy as the tiebreaker; this weighting is what makes the task non-trivial, since the 86.9% Failure imbalance rewards naive majority prediction under raw accuracy but not under Macro F_1 . Secondary metrics include per-class F_1 , precision, recall, and accuracy, together with two temporal metrics computed over positive-class clips (earliest detection time and false-negative rate per video) that reward early and consistent

detection. We also report video-level AUC-ROC, based on the maximum Failure-class probability across windows, as an additional secondary metric.

3 Method

3.1 Feature Extraction

Facial expression and acoustic features were extracted from the challenge datasets. To derive global facial expression features, we used the pretrained HSEmotion model to generate posterior probabilities for eight emotions (e.g., Anger, Contempt, Happiness) at 30 frames per second (fps) [10]. Furthermore, facial action unit (FAU) presence probabilities were extracted at 30 fps using an open-source facial expression analysis toolbox (Py-Feat)[3]. Py-Feat's head pose and emotion estimates were excluded due to missing head pose values (~1.8% of frames) and substantial overlap between the Py-Feat emotion estimates and HSEmotion features. Finally, 25 acoustic features from the eGeMAPSv02 set, including Mel-Frequency Cepstral Coefficients (MFCCs), jitter, shimmer, and loudness values, were extracted at 100 fps using the OpenSMILE toolkit [5, 6].

To retain local dynamic information before temporal alignment, first- and second-order temporal derivatives (Δ and $\Delta\Delta$ coefficients) were computed for the raw facial AUs, increasing the Py-Feat feature dimensionality to 60. All feature modalities were then temporally aligned by downsampling to 5 fps; we calculated mean values for acoustic features, and both mean and maximum values for facial features over each 0.2-second window. Since the aggregation process substantially increased feature dimensionality, we filtered out features with variance below 0.01 or collinearity above 0.95. This feature selection step removed 73 Py-Feat AU-derived, 8 HSEmotion, and 7 acoustic features, yielding a final feature set comprising 47 Py-Feat AU-derived, 8 HSEmotion, and 18 acoustic features. The end-to-end processing pipeline is visualised in Figure 1 of the supplementary materials.

3.2 Model Selection

To address the mismatch between the video-level ground truth labels and the challenge's requirement for window-level predictions, bystander affect detection was formulated as a multi-task learning setup that jointly learns frame and video representations. During inference, frame-level predictions were aggregated into window-level outputs via majority voting, following the challenge baseline.

Within this framework, two architectures were investigated: a cross-attention fusion transformer (symmetric and asymmetric configurations) and a bidirectional LSTM (BiLSTM).

3.2.1 Primary Approach: Cross-Attention Fusion Model. We adapted the cross-modal transformer of Venkatraman et al. [4] to operate over the multimodal features extracted in Section 3.1, rather than raw audio and video. A detailed visualisation of the adapted architecture is provided in Figure 2 of the supplementary materials.

Encoding. Acoustic and visual inputs were separated from the combined feature vector and optionally augmented with masked temporal differences. In the symmetric variant, both streams were embedded directly into a shared 128-dimensional space via identical linear encoders. In the asymmetric variant, audio was encoded with a multi-scale convolutional front-end, while video was encoded

with a smaller linear encoder; both streams were then projected into a common fusion space. In both variants, sinusoidal positional encodings were added per timestep before cross-modal attention.

Cross-modal transformer blocks. The encoded audio and visual sequences were passed through four stacked transformer blocks. Each block applied three sub-layers with residual connections and pre-layer normalisation: (1) within-modality self-attention, allowing each stream to build intra-modal temporal context; (2) bidirectional cross-attention, where the audio stream attends to the visual stream and vice versa, enabling cross-modality alignment and information exchange; and (3) a position-wise feed-forward sub-layer applied independently to each stream. The number of attention heads and the feed-forward expansion factor were optimised via hyperparameter tuning (more details in Section 3.3).

Gated fusion. After the final block, the audio $\mathbf{a}$ and visual $\mathbf{v}$ streams were combined via a learned sigmoid gate. Their concatenation $[\mathbf{a}; \mathbf{v}]$ was projected to compute a per-dimension scalar gate $g \in (0, 1)$, yielding $g \odot \mathbf{v} + (1 - g) \odot \mathbf{a}$. This representation was summed with a separate linear projection of the same concatenation and layer-normalised, producing a single fused sequence.

Multi-task prediction. Following the weakly supervised setup, the model produced frame- and video-level outputs from the fused sequence. A two-layer MLP classification head produced Frame-level logits at every timestep. For video-level prediction, a temporal attention pooling module computed a weighted sum over timesteps; the final video logit was the average of this pooled output and the masked mean of the frame-level logits. This dual aggregation ensured the video-level prediction was sensitive to both the peak discriminative moments and the overall sequence distribution.

3.2.2 Secondary Approach: Bidirectional LSTM. This architecture consisted of a BiLSTM layer followed by two task-specific heads. First, a frame-level head applied a linear projection to every hidden state, yielding per-timestep logits for frame-wise classification. Second, a video-level head used mean pooling followed by a linear projection to produce a video-level classification. In preliminary tests, the bidirectional model outperformed its unidirectional counterpart and was thus used in all subsequent experiments. An alternative approach using additive pooling attention to dynamically weight the contribution of each frame was also evaluated but omitted from the final system, as it failed to improve video-level performance.

3.3 Experimental Setup

To ensure a fair comparison between the cross-attention and BiLSTM models, both followed an identical experimental protocol. Input features were standardised using a *Z-score scaler* implemented via *scikit-learn*. Following the Track 1 requirements and with input features at 5 fps, window size and slide length were set to 2 seconds (10 frames) and 1 second (5 frames), respectively.

A subject-independent 5-fold cross-validation scheme was used for hyperparameter tuning to ensure model robustness. The Weights & Biases¹ platform was used to conduct random sweeps across various hyperparameters, including hidden size, learning rate, dropout, and batch size. We further evaluated different individual, pairwise,

and combined feature configurations. The optimal configurations were selected based on average validation video-level Macro F_1 -score (see Supplementary Table A1), retrained on the combined train/validation set, and evaluated on a held-out test set.

All models were implemented in *PyTorch* using the *Adam optimiser* and a *Cosine Annealing learning rate scheduler* featuring a 10-epoch linear warmup, where both the maximum and minimum learning rates were tuned. While initial experiments with the BiLSTM using video-level or frame-level loss functions struggled to capture either temporal nuances or global context, a multi-task formulation yielded substantial performance gains. Consequently, a multi-task hybrid loss function drove the training process:

$$\mathcal{L}_{\text{hybrid}} = \alpha \mathcal{L}_{\text{frame}} + (1 - \alpha) \mathcal{L}_{\text{video}} \quad (1)$$

To handle class imbalance, class weighting was applied to each loss component independently: video-level cross-entropy was weighted by the ground-truth class distribution, while frame-level loss used a tunable parameter to penalise false Control predictions. Early stopping and model checkpointing were used during cross-validation to prevent overfitting, with validation loss as the monitoring metric.

3.4 Generative AI Use Disclosure

Generative AI was used to assist with drafting, editing, and debugging the Python code for feature extraction, model training, and evaluation. We take full responsibility and accountability for the work and content of the paper.

4 Results

4.1 Model Performance

Tables 1 and 2 summarise model performance during 5-fold cross-validation and on the held-out test set, respectively. In both cases, cross-attention fusion outperformed the BiLSTM and the challenge baseline on video-level Macro F_1 and balanced accuracy. While the BiLSTM performance degraded from validation to final evaluation (F_1 : 0.719 to 0.636), the cross-attention models remained stable, indicating greater robustness to subject-independent test data. Despite this impressive performance, the cross-attention models achieved lower Control-class F_1 (~0.56) compared to the Failure-class (~0.94), due to strong class imbalance (86.9% Failure). However, the fact that the balanced accuracy of both cross-attention models is above chance confirms meaningful discrimination beyond majority-class bias.

4.2 Discriminative Feature Analysis

To locate discriminative signals in the training data, we computed Cohen's d effect sizes (Welch t-tests; Benjamini–Hochberg FDR) for all 73 features. Each feature was reduced to a single per-video value and compared between the 1,146 Failure and 173 Control videos. As shown in Table 3, Py-Feat action units showed the strongest individual effects, concentrated in smile-related features: lip-corner puller (AU12, $d = +0.53$), cheek raiser (AU06), and lip stretcher (AU20) were significantly elevated for Failure videos ($p_{\text{adj}} < 0.001$), while negative-reaction AUs (AU04, AU15) did not differ ($|d| < 0.12$, $p_{\text{adj}} > 0.4$). HSEmotion mirrored this trend, with significantly higher happiness scores and lower neutrality and anger scores on

¹<https://wandb.ai/>

Table 1: Model Performance on Track 1 using 5-fold cross-validation.

Level	Model	Macro F_1	Balanced Acc.	Accuracy	F_1 (+)	F_1 (-)	AUC-ROC
Video	BiLSTM	0.719	0.711	0.869	0.925	0.499	0.753
	Cross-Attention (Symmetric)	0.752	0.733	0.897	0.941	0.563	0.821
	Cross-Attention (Asymmetric)	0.752	0.739	0.894	0.940	0.565	0.860
Window	BiLSTM	0.732	0.763	0.856	0.891	0.573	0.802
	Cross-Attention (Symmetric)	0.805	0.794	0.895	0.936	0.674	0.876
	Cross-Attention (Asymmetric)	0.811	0.804	0.894	0.936	0.686	0.895

Note: This table reports the average scores across the 5-folds for each window-and video-level metric. (+) denotes the Failure class (86.9%) and (-) denotes the Control class.

Table 2: Model Performance on the Track 1 held-out test set.

Model	Macro F_1 ($\uparrow$)	Balanced Acc. ($\uparrow$)
BadNetCNN (Official Baseline)	0.502	0.504
BiLSTM	0.636	0.610
Cross-Attention (Symmetric)	0.767	0.746
Cross-Attention (Asymmetric)	0.744	0.768

Note: This table reports the primary video-level metrics for ERR@HRI 3.0.

Table 3: Most discriminative features per modality on Track 1, ranked by effect size $|d|$.

Feature	d	Higher in	p_{adj}
Py-Feat action units			
AU12 lip-corner puller	+0.53	Failure	< 0.001
AU06 cheek raiser	+0.50	Failure	< 0.001
AU20 lip stretcher	+0.46	Failure	< 0.001
HSEmotion			
Neutral	-0.40	Control	< 0.001
Happiness	+0.39	Failure	< 0.001
Anger	-0.39	Control	< 0.001
Acoustic (eGeMAPS)			
F2 frequency	-0.07	Control	0.80
MFCC3	+0.06	Failure	0.85
MFCC2	-0.06	Control	0.86

Note: $d > 0$ indicates higher values on Failure videos. p_{adj} computed via Welch's t-test with Benjamini–Hochberg FDR correction over 73 features ($\alpha = 0.05$).

Failure videos. These results suggest that robot failures elicited amused reactions rather than distress or anger from bystanders.

The eGeMAPS acoustic features separated the classes poorly ($\max |d| = 0.07$, $p_{\text{adj}} = 0.80$). However, combining the acoustic and facial streams consistently yielded optimal BiLSTM performance during hyperparameter tuning, suggesting that the acoustic features provide complementary information not captured by facial features alone. We therefore retained both streams for our submitted models.

5 Discussion and Conclusion

This study evaluated a cross-attention fusion architecture against a BiLSTM with simple concatenation for Track 1 of the ERR@HRI 3.0 challenge: Bystander Affect Detection. Because these models shared the same features, alignment, and supervision, the performance

gap between them can be attributed to cross-attention effectively modelling the relationships between the facial and acoustic inputs. Notably, while cross-attention performed similarly on the validation and test sets, BiLSTM performance degraded sharply, highlighting cross-attention's superior robustness to inter-participant distribution shift. However, both approaches struggled to predict the Control class, a direct consequence of the (~87%) imbalance. The fact that the balanced accuracy for all three submitted models substantially exceeded that of a naive majority predictor confirms that they discriminate between the two classes rather than relying on the prior. Further analysis revealed that temporally escalating smile signals were the strongest discriminative cues, challenging the assumption that bystanders react to errors with negative expressions and underscoring the role of temporal dynamics.

Several factors should be considered when interpreting our results. First, the small Control test set (43 of 326 videos) makes Macro F_1 sensitive to minor prediction shifts, limiting how confidently the results can be interpreted. Second, using pre-extracted HSEmotion, Py-Feat, and OpenSMILE features leaves the pipeline vulnerable to variations in background noise, lighting, and camera angles, leading to feature-extraction errors that propagate into the classifier.

These limitations suggest several directions for future work. The class imbalance could be tackled during training through focal loss, class-balanced sampling, or minority-class data, which would encourage the model to attend to the Control class. Feature reliability could be improved through targeted pre-processing—such as filtering out irrelevant ambient noise from the audio and applying illumination normalisation to the video—before feature extraction. Finally, extending the approach to Track 2 would test further whether cross-attention fusion generalises to the anticipatory-outcome setting.

Acknowledgments

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT), their applications funded by UK Research and Innovation [grant number EP/S023062/1], institutional funding from the University of Sheffield (PhD studentship support, 236203181 and 255772965), and Innovate UK (grant number 10089807 for the European Union's Horizon Europe research and innovation programme under Grant Agreement n° 101120727-PRIMI) and UKRI (EPSRC) Grant EP/X018733/1 (ALDENS). The views expressed are those of the authors and do not necessarily reflect those of the funders.

References

- [1] Alexandra Bremers, Maria Teresa Parreira, Xuanyu Fang, Natalie Friedman, Adolfo Ramirez-Aristizabal, Alexandria Pabst, Mirjana Spasojevic, Michael Kuniavsky, and Wendy Ju. 2023. The Bystander Affect Detection (BAD) Dataset for Failure Detection in HRI. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 11443–11450. doi:10.1109/IROS55552.2023.10342442
- [2] Shiye Cao, Maia Stiber, Amama Mahmood, Maria Teresa Parreira, Wendy Ju, Micol Spitale, Hatice Gunes, and Chien-Ming Huang. 2025. ERR@HRI 2.0 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Conversations. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14130–14135.
- [3] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. 2023. Py-Feat: Python Facial Expression Analysis Toolbox. *Affective Science* 4, 4 (2023), 781–796.
- [4] Joe Dhanith P R, Shravan Venkatraman, Vigya Sharma, and Santhosh Malarvannan. 2026. Multimodal emotion recognition using audio-video transformer fusion with cross attention. *arXiv preprint arXiv:2407.18552* (2026).
- [5] 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. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.
- [6] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*. 1459–1462. doi:10.1145/1873951.1874246
- [7] Xun Jiang, Shuangle Li, Chong Liu, and Xing Xu. 2025. Multimodal Time Series Alignment for Error Detection in Human Robot Interactions. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14143–14149.
- [8] Maria Teresa Parreira, Micol Spitale, Maia Stiber, Shiye Cao, Amama Mahmood, Chien-Ming Huang, Hatice Gunes, and Wendy Ju. 2026. ERR@HRI 3.0 Challenge: Multimodal Detection of Errors and Anticipation in Human-Robot Interactions. In *Proceedings of the 28th ACM International Conference on Multimodal Interaction (ICMI '26)*. ACM.
- [9] Rutherford Agbeshi Patamia, Ha Pham Thien Dinh, Ming Liu, and Akansel Cosgun. 2025. Beyond Technical Failures: Multimodal Time-Series Modelling for Detecting Social Breakdowns and User Repair Attempts in Human-Robot Interaction. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14136–14142.
- [10] Andrey V Savchenko. 2022. HSEmotion: High-speed emotion recognition library. *Software Impacts* 14 (2022), 100433.
- [11] Micol Spitale, Maria Teresa Parreira, Maia Stiber, Minja Axelsson, Neval Kara, Garima Kankariya, Chien-Ming Huang, Malte Jung, Wendy Ju, and Hatice Gunes. 2024. ERR@HRI 2024 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Interactions. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 652–656.
- [12] Maia Stiber, Russell Taylor, and Chien-Ming Huang. 2022. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 676–683.
- [13] Lennart Wachowiak, Peter Tisnikar, Andrew Coles, Gerard Canal, and Oya Celiktutan. 2024. A time series classification pipeline for detecting interaction ruptures in HRI based on user reactions. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 657–665.