

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 Supplemental Material

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>

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

A.1 Feature Extraction Pipeline

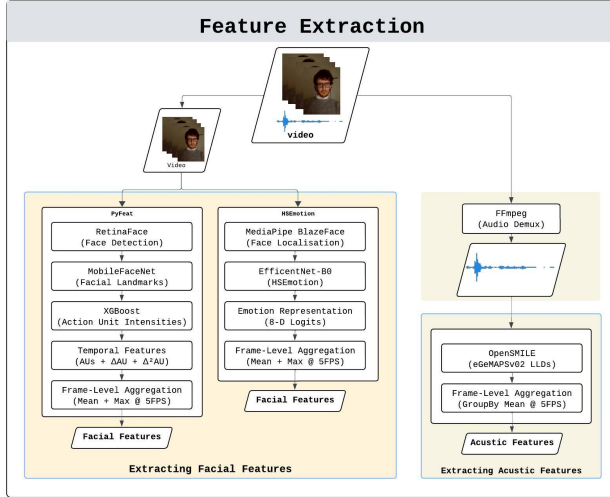


Figure 1: Feature extraction pipeline used for ERR@HRI 3.0 Track 1. Py-Feat facial action units (FAU) and emotion logits (HSEmotion) were extracted at 30 fps and pooled to 5 fps; acoustic eGeMAPS descriptors were extracted at 100 fps and mean-aggregated to 5 fps. All modalities shared a common temporal grid before variance/collinearity filtering.

A.2 Cross-Attention Architecture

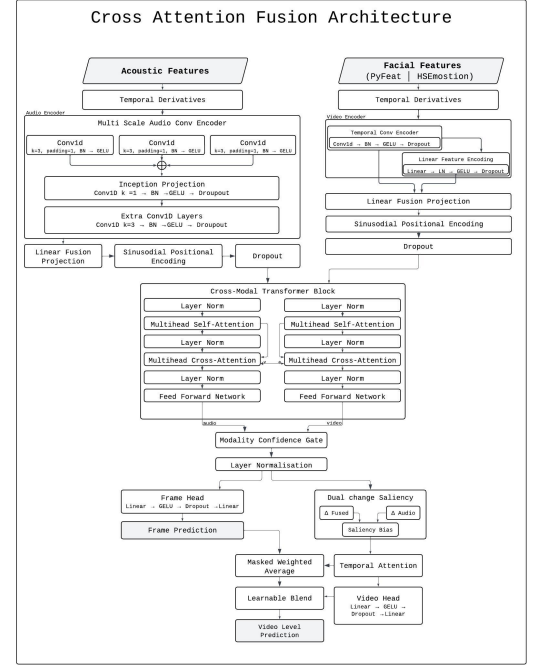


Figure 2: Cross-attention fusion architecture (v2). Acoustic and facial feature streams are encoded asymmetrically, interact through stacked cross-modal transformer blocks, and are fused via a modality-confidence gate before frame-level and video-level prediction heads.

A.3 Hyperparameter Tuning Results

Table A.1: Optimal Hyperparameter Configurations for the BiLSTM and Cross-Attention Models

Hyperparameter	BiLSTM	Cross-Attention
Input Features	HSEmotion + Acoustic	HSEmotion + Acoustic
<i>Model Architecture</i>		
Hidden Size	128	128
N Layers	1	4
Convolution Layers	–	1
N Attention Heads	–	8
FF Multiplier	–	2
<i>Optimisation & Training</i>		
Batch Size	64	16
Epochs	50	30
Warmup Epochs	10	10
LR	6.25×10^{-5}	3.91×10^{-4}
Minimum LR Ratio	6.99×10^{-3}	6.69×10^{-4}
<i>Regularisation & Loss</i>		
Dropout	0.295	0.162
Weight Decay	1.00×10^{-5}	8.51×10^{-5}
Loss weight (α)	0.400	0.427
Frame Loss Weight	4.31	1.51
Focal Gamma	–	1.13
Label Smoothing	–	0.084

Note: FF Multiplier = Feed Forward Multiplier, LR = Learning Rate.