Two-Stage Multimodal Framework for Emotion Mimicry Intensity Prediction

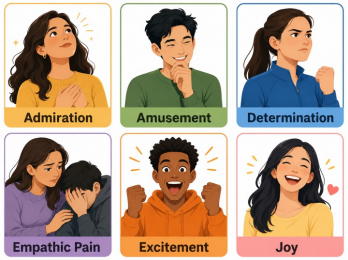
Dinithi Dissanayake[1], Shaveen Silva[1,2], Ovindu Atukorala[1,2], Prasanth Sasikumar[1], Suranga Nanayakkara[1]
[1]National University of Singapore, [2]University of Moratuwa

CVPR JUNE 3-7, 2026 DENVER COLORADO

EMI Challenge[1]

Goal: Predict six continuous emotional mimicry intensity dimensions from in-the-wild multimodal video clips.

Target emotions:



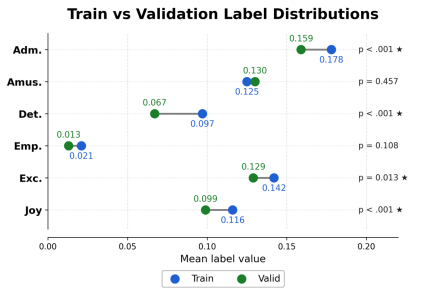
Key challenges:

- Long-tailed sequence lengths
- Sparse and imbalanced labels
- Missing or incomplete text modality
- Noisy in-the-wild audio-visual signals

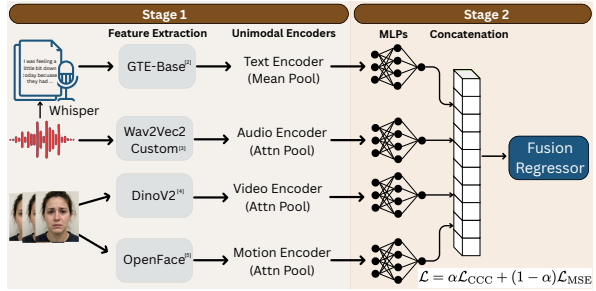
Sys.	Mods.	Fusion	Split	Val.	Test
F1	T+A+V	MLP	2:1	0.4373	0.506
F2	T+A+V+M	MLP	2:1	0.4386	0.506
F3	T+A+V	MLP	4:1	0.4717	0.570
F4	T+A+V+M	MLP	4:1	0.4722	0.562
F5	T+A+V	MLP (frz.)	2:1	0.4400	0.491

Our submission: Fusion performance across modality combinations and data splits.

Methods and Materials



Dataset split shows sparse label distributions.

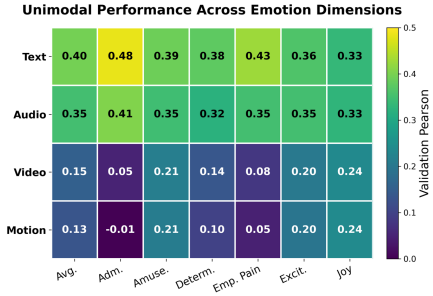


Our two-stage framework: unimodal feature learning followed by multimodal fusion.

Results

Modality	Input / Feature	Model	Val. Pearson
Visual	ViT face embed.	GRU	0.1084
	DINOv2 full-frame embed.	GRU	0.1343
	DINOv2 full-frame embed.	LSTM	0.1260
	ViT face embed.	LSTM	0.1134
	DINOv2 face embed.	GRU	0.1512
	DINOv2 face embed.	Attention pooling	**0.1535**
Audio	wav2vec2 embed.	Linear projection	0.1428
	wav2vec2 embed.	LSTM	0.3088
	Custom wav2vec2 embed.	Mean pooling + MLP	0.3500
	Custom wav2vec2 embed.	Mean+std pooling + MLP	0.3600
	Custom wav2vec2 embed.	Attention pooling	0.3600
	Custom wav2vec2 embed.	3-layer BiLSTM + attention	**0.3650**
Text	GTE embed.	LSTM	0.4070
	GTE embed.	MLP	**0.4076**
	BERT-Large embed. (CLS)	MLP	0.3400

Validation performance of individual modality encoders.



Modalities provide complementary cues while also reinforcing common affective patterns.

Discussion

- **Text and audio** were the **strongest unimodal** predictors, showing the importance of semantic and vocal cues.
- Visual and motion features were weaker alone but still provided **complementary** affective information.
- The **4:1 split improved performance**, suggesting fusion benefits from more supervised training data.
- Motion gave only modest gains, indicating that **temporal facial dynamics need better modeling**.
- **Possible future improvements**
  - Use pretrained feature extractors specialized for affective modeling.
  - Handle missing modalities more explicitly using learned missing-modality tokens.
  - Add auxiliary supervision to preserve strong unimodal signals.

Conclusions

- We proposed a two-stage multimodal framework for EMI prediction using text, audio, vision, and motion.
- Multimodal fusion improved performance over unimodal models.
- Our best system achieved **0.57 test Pearson** and ranked **third** in the EMI challenge.
- Future work can explore stronger motion representations and more adaptive fusion methods.

Contact

Dinithi Dissanayake
National University of Singapore
Email: dinithid@u.nus.edu
Website: https://dinithidissanayake.com/

References

1. Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Ste- fanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Peder- soli, Simon Bacon, Alice Baird, Chris Gagne, et al. Ad- vancements in affective and behavior analysis: The 8th abaw workshop and competition. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5572– 5583, 2025.
2. Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281, 2023
3. Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460, 2020
4. Maxime Oquab, Timoth´ee Darcet, Th´eo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision.arXiv preprint arXiv:2304.07193, 2023.
5. Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In 2016 IEEE winter conference on applications of computer vision (WACV), pages 1–10. IEEE, 2016