Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD38408

A Reliability Sensitive Theory of Multimodal Emotion Recognition

Yihang Chen

School of Electrical and Electronic Engineering, The University of Manchester, Manchester, UK
yihang.chen-5@student.manchester.ac.uk

Abstract. Current research on multimodal emotion recognition seeks to analyse the relationships among facial, vocal, and bodily expressions in order to recognise different emotions. Although additional data streams are often introduced, they cannot be treated as independent evidence. This conceptual study aims to understand the reliability of multimodal inferences over unimodal classifications. This study covers the theory and practice of affective computing, the study of different modalities, fusion taxonomies and benchmarking. Depending on the statistical dependencies that they learn, fusion can occur in early, intermediate, late or adaptive form. The synthesis suggests that the advantage of multimodal systems is determined by four factors: informational complementarity, partially independent errors, temporally valid alignment, and dynamic reliability estimation. For these factors, fusion of different types of data is a good alternative. Otherwise, it is likely to increase the recording of social displays, bias of annotations, or dependency on a dominant modality. For this reason, a reliability-sensitive framework has been proposed, in which recognition is driven by uncertainty management rather than feature accumulation, while preserving the modality-specific nature of different types of evidence. The resulting account explains both the recurrent gains reported in benchmark studies and their contraction under naturalistic expression, distribution shift, cultural mismatch, strategic display, and systematic missingness. Multimodal systems become defensible not by sensing more behaviour alone, but by estimating when each observation should change, confirm, or weaken an affective inference.

Keywords: multimodal emotion recognition, affective computing, cross-modal fusion, reliability, uncertainty

1. Introduction

Emotion recognition aims to infer affective states from observable behaviour. Affective computing established this problem as one of sensing, representing, and responding to affect, but behaviour is not a transparent measurement of inner experience [1]. A smile may signify many feelings and gestures. Embarrassment and politeness can also be expressed with a smile. The same applies to shouting, which may express anger or excitement, or to struggling to be heard in a noisy environment. Single-channel systems are liable to the failings of the channel used. Multimodal systems aim to address uncertainty by incorporating Facial, vocal, text and bodily evidence.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD38408

Research has moved from feature concatenation to representation learning, temporal attention, and adaptive fusion [2]. The architectural literature describes multimodal advantage more confidently than its conceptual basis warrants. Additional modalities may repeat the same information, share the same bias, arrive at incompatible times, or disappear in deployment. Average benchmark accuracy cannot show whether a model integrated complementary evidence or learned a stable shortcut, and diagnostic studies have found substantial dependence on a single modality [3]. An inquiry of this nature looks at the circumstances in which the presence of facial, vocal, and bodily cues facilitates better and more complete inferences than any given cue on its own. It also looks at how some fusion protocols manage to keep the uncertainty that remains when different modalities cease to agree or degrade. Using critical conceptual synthesis, I attempt to develop a reliability-sensitive account. Prior work on affective data, cross-cultural variations, multiple modal learning, and dataset robustness is assessed based on four conditions: informational complementarity, error decoupling, temporal alignment, and dynamic reliability estimation. Then, the comparison and evaluation of fusion protocols are based on the same four conditions. Potential implications regarding the design of high-stake datasets are then discussed along with other issues of interest like calibration and missing modalities. Finally, this structure shows how the choice of synthetic models can be connected to the assumptions that justify the trusting of such models and can also articulate standards of future model development.

2. Conditions and mechanisms of reliable integration

This section describes the necessary conditions for modalities to enhance emotion inference. It distinguishes complementary information from cues that are redundant or conflicting, then analyses temporal alignment and interactional context. The subsequent analysis compares fusion architectures based on how they model dependence and how they manage inputs that are either wholly or partially unavailable. The final section sets out goals for balancing confidence and model abstention.

2.1. Complementarity redundancy and conflict

Complementarity provides the most direct route to multimodal gain. A meta-analysis of 90 systems found that 85% of multimodal systems exceeded their strongest unimodal comparator, although gains were smaller for natural than acted expressions [4]. If each modality has unique information about the class, their combined observation will have less ambiguity regarding the answer. For example, quick, intense speech accompanied by a furrowed brow makes anger or concentration more likely than neutrality. One will be more likely to interpret someone as sad vs. angry/neutral, if that person has a posture slouched over combined with soft neutral speech. What matters is whether the second modality provides case-specific information. This second modality will be helpful if it alters the range of emotions that are likely to be true, after the first modality is observed. Just combining both modalities doesn't provide context. Jointly de-correlating errors makes a system less vulnerable to failure. If face, voice, and body cues do not agree, it is preferred to discard the prediction, even if it is locally efficient. For face, voice, and body cues each to be separately efficient, the joint observation must not be interpreted as independent evidence.

Cross-modal conflict has evidential value when the system can distinguish meaningful incongruence from technical failure. Positive words delivered with flat prosody and a frown may signal sarcasm, resignation, interpersonal tension, or a delayed response. Forced agreement discards those possibilities, while a single fused vector can conceal which channel generated the conflict. The critique of emotion inference from isolated facial movement becomes more consequential when an

DOI: 10.54254/2755-2721/2026.LD38408

architecture converts ambiguous signals into an apparently coherent class [5]. Modality-specific predictions should remain inspectable until reliability and temporal relations have been assessed.

2.2. Temporal alignment and interactional context

Temporal alignment determines whether observations refer to the same expressive event. Acoustic frames, facial actions, lexical units, and gestures occur at different rates and can lead or follow one another. Laughter may begin vocally before a smile reaches its peak, while a posture change may precede a verbal disclosure. IEMOCAP provides synchronised speech, video, and motion capture because expressive interaction cannot be reduced to independent snapshots [6]. Alignment based on file timestamps is inadequate when the relevant unit is an interval, conversational turn, or response sequence.Interactional context supplies evidence that local features omit. MELD shows how speaker history and preceding turns change the interpretation of a current utterance [7]. Emotion-shift modelling and multimodal graph architectures make this dependence explicit by relating the current expression to changes across conversational turns [8, 9]. The same short response can have different meanings depending on what was said before, or who said it. Contextual meaning must not be confused with the historical context which can include other factors unrelated to the emotion, such as the topic, identity, or production of the artefacts. A model that uses context should investigate whether contextual factors enhance transfer out of the corpus in which they were observed.

2.3. Fusion architecture and dynamic reliability

Fusion architectures make different assumptions about dependence. In early fusion, low-level features or their summaries are combined to produce predictors. These architectures can model closeness, but are fragile to variations in scale, sampling, and missing values. In intermediate fusion, encoding is performed for each of the modalities. Information is then passed across modalities in a shared latent space. Layer-wise fusion can retain modality-independent representations while aggregating cross-modal information [10]. High model capacity also makes intermediate models vulnerable to dataset-specific associations whose evidential relevance is hard to verify.

Late fusion combines probabilities or scores of separate modality decisions. It is modular, so it can support missing channels, reliability weighting and the examination of disagreement. However, interactions that unimodal encoders discard cannot be recovered at the decision stage. Adaptive hybrids connect these levels by preserving private representations, learning cross-modal relations and gaining contributions based on quality estimation. The advantage of these approaches is that they are highly dependent on the gate. In some cases, the gate introduces a second prediction problem that must be corrected. From the results in Table 1, there is no clear winner in terms of fusion stage in all conditions. Early fusion can handle highly synchronised streams, but it is not ideal for missing inputs. Intermediate fusion retains flexible temporal dependence, but it makes shortcut detection more challenging. Late fusion preserves disagreement and modularity but sacrifices fine-grained interactions. Adaptive hybrids give users the most control over contribution weights, but weighing contributions that are too inconsistent can intensify modality dominance. The most appropriate architecture remains unpredictable and should not be based solely on an uninformed abstract framework of representational depth.

DOI: 10.54254/2755-2721/2026.LD38408

Table 1. Theoretical comparison of fusion strategies

Strategy	Interaction capacity	Missing-modality tolerance	Main vulnerability	Best fit
Early	High	Low without imputation	Alignment and scale error	Synchronous streams
Intermediate	High	Medium with masked encoders	Opaque interactions and shortcuts	Long-range dependence
Late	Low	High with valid unimodal models	Lost feature interactions	Intermittent streams
Adaptive hybrid	Selective	High if gating is calibrated	Gate failure and complexity	Variable signal quality

2.4. Reliability sensitive decision criteria

Multimodal superiority requires four jointly relevant conditions. The modalities must add non-identical information, their errors must be sufficiently decoupled, observations must be aligned at the timescale of the behaviour, and reliability must be estimated during inference. Aggregate benchmark gains can coexist with strong dependence on a single modality [3]. Missing-modality research shows that absent channels can alter semantic predictions and sharply reduce performance [11, 12]. Modality dropout, quality-aware gates, calibrated ensembles, and explicit missingness indicators address different failure modes, but each should be tested against the condition it claims to satisfy.

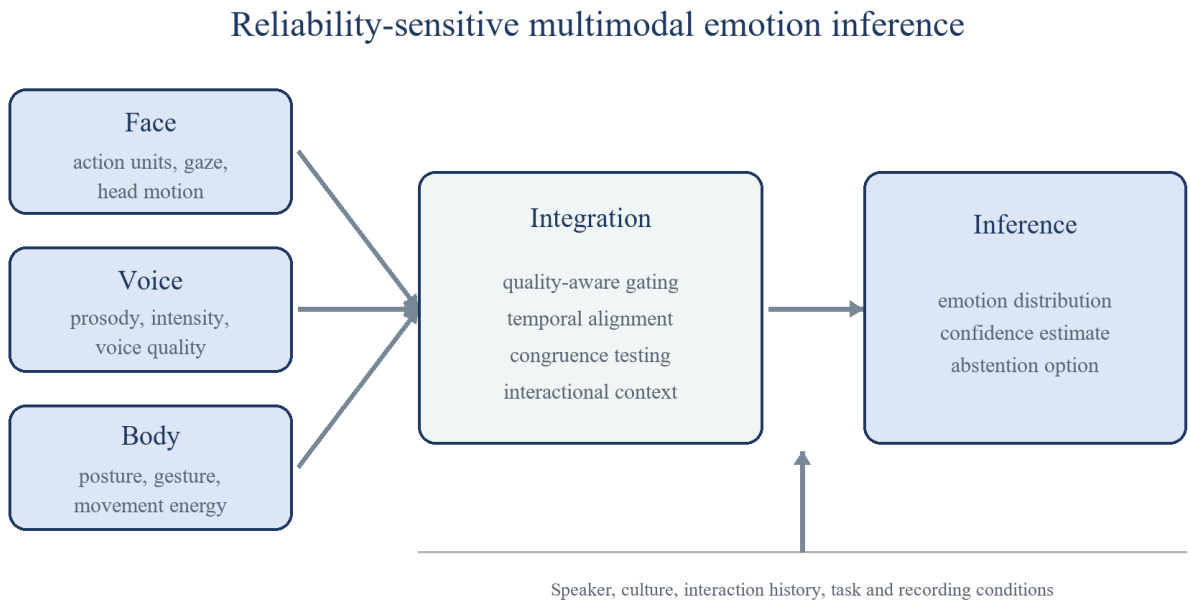


Figure 1. Reliability-sensitive integration of multimodal emotion cues

The decision rule must synthesize both uncertainty and error cost. When all modalities are weak, agreement may reflect a shared artefact rather than strong evidence. When reliable modalities conflict, the posterior should remain dispersed. Thus, confidence must be evaluated against the observed error, and the model should abstain from making a decision when the selective risk

DOI: 10.54254/2755-2721/2026.LD38408

exceeds a threshold. As represented in Figure 1, this framework considers the quality, alignment, and congruence of modalities and the context of the interaction as intervening variables in representation learning. Modality-specific encoders preserve evidence to audit how the output distribution was formed. For probabilistic outputs, one source of uncertainty is ambiguity in the behaviour, while the other reflects the model's insufficient knowledge. The first type can be present even with observations of well-documented behaviour, as multiple interpretations may still exist. The second can be expected to increase in the presence of novel speakers, cultures, and devices or in the presence of inconsistencies. Collapsing both types of uncertainty as the maximum softmax score obscures ambiguity and weakens decisions regarding the collection and abstention of data. With this, reliability-sensitive integration of disparate, calibrated distributions informs the decision through distinct, inspectable pathways.

3. Discussion

Benchmark design limits the scope of performance claims. IEMOCAP comprises about 12 hours of recordings from 10 actors in controlled dyadic sessions, while MELD contains about 13,000 utterances drawn from a television programme [6, 7]. Although the two datasets are comparable, neither captures the full range of spontaneous interaction across languages and cultures. Acted expression can help show cross-modal agreement. Research dialogue contains production factors that are typically not found in regular settings. Because of this, subject-independent and cross-corpus tests show stronger evidence for generalisation than randomly splitting utterances. Another limitation is shown when annotation sets a second boundary. Emotion categories are essentially observer assessments of behaviour and are not real measurements of a mental state. Disagreement can be said to reflect uncertainty, mixed emotion, cultural interpretation, or an inadequate label set. Majority voting removes all this and encourages overconfident outputs. Probability distributions, ordinal dimensions such as valence and arousal, and explicit annotator models keep more of the target uncertainty. Labels impact the architecture, and because of this, data collection and fusion designs are strongly coupled.

Privacy and institutional power constrain legitimate use even when technical accuracy is adequate. Audio, facial video, and body motion can reveal identity, health, disability, and social behaviour beyond the intended label. Collection should minimise retained raw data, separate identity from affect features, document access, and permit withdrawal where feasible. High-stakes deployment needs a contestable process because an affect estimate may be wrong without visible sensor failure. Regulation (EU) 2024/1689 prohibits emotion inference in workplaces and educational institutions except for medical or safety purposes [13].

Future models must consider modality, quality and missingness in model design as primary factors. Prior to fusion quality-aware gates must assess fidelity and domain. Missingness must distinguish an inactive stream from a neutral, emotionless observation. Compared to concatenation, privately held representations provide no reliability adjustment. Evaluation must include calibration error and negative log-likelihood; selective risk assessment; cross-speaker and cross-corpus transfer; measures of systematic masking and graded corruption; and assessment of subgroup performance and abstention frequency. Such tests expose dominant-modality shortcuts and distributional fragility that mean accuracy can conceal [3, 11, 12]. Reporting should distinguish performance with all streams present from performance under each controlled absence pattern. Interpretability can identify which observations altered a prediction, but attribution cannot convert ambiguous behaviour into ground truth; claims should remain probabilistic and conditional [5].

DOI: 10.54254/2755-2721/2026.LD38408

4. Conclusion

The research question is concerned with what circumstances enable the integration of facial, vocal, and bodily cues to allow for affective inferences that are greater in both scale and accuracy than regardless of cue presence. The analysis outlines four conditions. The respective modalities must supply complementary information, their errors must be sufficiently independent, observations must be synchronised to the same behavioral event, and the system must evaluate the changing conditions to determine reliability. Inconsistency can signal ambiguity, regulation, or interaction rather than mere noise. Inconsistency does not operate automatically. Agreement may reinforce a bias, and conflict may occur due to timing or sensor failure rather than affective complexity.

Fusion strategies control how identifiable the operations are. Early fusion captures neighbouring feature interactions, but is susceptible to heterogeneous scales and missing values. Intermediate fusion adds depth to the representational semantics and the temporal semantics at the cost of model interpretability. Late fusion preserves modularity and disagreement but loses interactions that unimodal encoders cannot recover. Hybrid models with adaptive fusions allow the most control of the operations when their gates are calibrated and validated. The correct architecture is, therefore, the one that matches the assumptions of the deployment situation, anticipated gaps, cost of sensing, and the expected impact of erroneous decisions.

This paper suffers from a conceptual method and synthesis benchmark literature limitations. It does not examine causality of architecture, and gains reported cannot be generalized to other, unstudied populations or institutions. Future research should combine cross-speaker, cross-corpus assessment with controlled corruption, missing-modality tests, calibration, subgroup analysis, and selective prediction. Model uncertainty should be maintained, the possible role of each modality should be disclosed, and the system should abstain when the evidence is insufficient. Data minimization, purpose limitation, and contestability remain important for affective classifications where they influence access, assessment, or employment. Empirical tests of the four conditions should examine not only whether fusion is improved with respect to a score, but which source of uncertainty was changed and if the improvement is also observed with respect to distribution shift. Because of these considerations, multimodality is best viewed as a conditional approach for uncertain behavioral evidence, rather than approaching it as a certain means for improving knowledge if more behavior is sensed.

References

[1] Picard, R. W. (1997). *Affective computing*. MIT Press. Cambridge, MA. https://mitpress.mit.edu/9780262661157/affective-computing/.

[2] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.

[3] Hazarika, D., Li, Y., Cheng, B., Zhao, S., Zimmermann, R., & Poria, S. (2022). Analyzing modality robustness in multimodal sentiment analysis. *Proceedings of NAACL 2022*, 685–696. Seattle. https://doi.org/10.18653/v1/2022.naacl-main.50.

[4] D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys, 47*(3), Article 43, 1–36. https://doi.org/10.1145/2682899.

[5] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. https://doi.org/10.1177/1529100619832930.

[6] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation,*

DOI: 10.54254/2755-2721/2026.LD38408

42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[7] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL 2019* (pp. 527–536). Association for Computational Linguistics. Florence. https://doi.org/10.18653/v1/P19-1050

[8] Bansal, K., Agarwal, H., Joshi, A., & Modi, A. (2022). Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models* (pp. 44–56). https://aclanthology.org/2022.mmmpie-1.6/

[9] Xu, S., Jia, Y., Niu, C., & Zan, H. (2022). MMDAG: Multimodal directed acyclic graph network for emotion recognition in conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6802–6807). European Language Resources Association. Marseille.https://aclanthology.org/2022.lrec-1.733/

[10] Sun, J., Han, S., Ruan, Y.-P., Zhang, X., Zheng, S.-K., Liu, Y., Huang, Y., & Li, T. (2023). Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 1*, 658–670. https://doi.org/10.18653/v1/2023.acl-long.39

[11] Lin, R., & Hu, H. (2023). MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics, 11*, 1686–1702. https://doi.org/10.1162/tacl_a_00628

[12] Zeng, J., Zhou, J., & Liu, T. (2022). Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2924–2934). Association for Computational Linguistics. Abu Dhabi.https://doi.org/10.18653/v1/2022.emnlp-main.189

[13] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. http://data.europa.eu/eli/reg/2024/1689/oj