

International Journal of AI and Advanced Computing

ISSN: 3141-3161

journal homepage: <https://academiansedu.org/ijaac/>

Research Article

A Unified Uncertainty-Aware and Imbalance-Resilient Framework for Robust Multimodal Emotion Recognition

Qasim Mahdi Haref¹, *, Murteza Hanoon Tuama¹, Wahhab Muslim Mashloosh¹¹ Computer Techniques Engineering, Imam Alkadhim University College (IKC), Baghdad, IraqCorresponding authors email: qasimmahdi@iku.edu.iq

ARTICLE INFO

Article history:

Received: 08 July 2026

Received in revised form:
18 August 2026Accepted: 12 September
2026Available online: 30
September 2026

Keywords:

Multimodal Emotion
Recognition;
Uncertainty-Aware
Learning; Class
Imbalance Mitigation;
Adaptive Fusion;
Contrastive
Representation Learning;
Affective Computing

ABSTRACT

The multimodal emotion recognition has become a key field of research because of its importance in human-computer interaction, affective computing and intelligent decision support systems. But these techniques are generally hampered by modality imbalance, class distribution imbalance and propagation of uncertainty across heterogeneous feature space. To solve these problems, the proposed study provides a unified and modular multimodal learning framework based on four complementary enhancement mechanisms to boost the robustness, generalization, and interpretability. The proposed framework consists of (i) an uncertainty-aware adaptive gating mechanism to control the contributions from the different modality, (ii) a multi-sample dropout strategy for better regularization and to prevent overfitting, (iii) a contrastive representation alignment module to ensure the consistency of the different modalities, and (iv) a logit-adjusted learning scheme to deal with the issue of class imbalance. All are organized as separate components, which allows for flexible activation and for systematic ablation analysis. The proposed framework consistently outperforms baseline emotion models and all individual variants of the proposed enhancement on the weighted F1-score, macro F1-score, and overall classification accuracy on extensive experiments conducted on a large-scale multimodal emotion dataset. The ablation results also validate the contribution of uncertainty modeling and imbalance-aware learning, as they yield the best performance. Interestingly, the full configuration works well on CPU-based environments and provides the best generalization performance. The work presented in this thesis makes a significant contribution to the scalable, interpretable, and practically deployable multimodal emotion recognition framework, and provides insights into future research for developing adaptive multimodal learning and powerful affective computing systems.

Keywords:

DOI:
[10.36099/ijaac.v1i3.002](https://doi.org/10.36099/ijaac.v1i3.002)

1. Introduction

In recent years, sentiment analysis and emotion detection have come into the focus of modern intelligent systems because of their increasing importance in the human-computer interaction, affective computing, and decision support by data fields. The initial studies in this area were mainly restricted to coarse sentiment classification of the basic sentiment categories (positive, negative and neutral), where "polarity" refers to positive or negative sentiment. These simplifications, however, do not reflect the nuances and dynamics of human emotions, especially in practical, real-world scenarios (Tuama et al., 2023). This has led to the recent focus on the need to go beyond polarity to emotion recognition and sentiment analysis

frameworks with a focus on impact (Tuama et al., 2023; Y. Wu et al., 2025). Concurrently, sentiment analysis and emotion recognition studies have been dramatically changed by the surge of multimodal data, which includes text, speech, and visual data. Multimodal emotion recognition (MER) seeks to take advantage of complementary information from heterogeneous modalities to gain more powerful and human-like emotional understanding (Y. Wu et al., 2025; C.-Y. Wu et al., 2025). Despite significant progress, current multimodal methods still suffer from common issues of modality imbalance, unreliable or noisy signals and inconsistent cross-modal representations (Hamidi et al., 2025; Devarajan, 2025).

They can cause unstable fusion behavior and poor generalization, especially when modalities are given the same importance as models assume, even if the reliability of the modalities differs (Chen, 2025). Class imbalance is another crucial challenge presented in existing emotion recognition systems, where a small percentage of the emotions are expressed by a large proportion of the data. This imbalance leads to biases in model predictions and decreases the interpretability and fairness in affective computing applications (Al-Azani et al., 2025; Zhang et al., 2025). Recent learning strategies for improving minority emotion recognition have shown good performance gains such as logit adjustment and prior-based optimization, but they tend to be explored individually and not integrated with the wider multimodal fusion strategies (Mandal & Li, 2025). Furthermore, the main problem for high-dimensional multimodal feature spaces is overfitting, particularly when there is limited computational resources, CPU based deployment environment, etc. (Lee et al., 2025). Recently, the representation robustness has been further improved by focusing on contrastive learning to align the embeddings for different modalities and minimize semantic inconsistencies between modalities (Li et al., 2024; Farhadipour et al., 2025). There is not much research on contrastive alignment and its combination with uncertainty modeling and imbalance-aware learning. Existing research focuses on these enhancement strategies individually and provides disjointed solutions that are hard to adapt, interpret, and systematically examine in various datasets or deployment contexts (C.-Y. Wu et al., 2025; Kumar & Sharma, 2025). Inspired by these gaps, this study follows earlier research where the more general applications and societal implications of sentiment analysis and emotion detection beyond polarity-based modelling were noted (Tuama et al., 2023). We suggest a common and modular multimodal learning model that incorporates uncertainty-aware fusion, robust regularization, contrastive representation alignment, and imbalance-aware optimization in a single architecture. All of the components are independent modules and can be activated independently, which allows for careful ablation analysis. This paper has validated that the combination of these mechanisms has a consistent ability to enhance the weighted F1-score, macro F1-score, and overall classification accuracy in CPU-based environments, with a large-scale multimodal emotion dataset (Y. Wu et al., 2025; Al-Azani et al., 2025; Ahmed & Rahman, 2025; Liu, 2025). Therefore, the proposed framework provides a scalable, interpretable and practically deployable approach for next generation affective computing systems.

In this research, the following points of significance are made:

1. **Emotion Recognition Based on Modular Multi Modal Architecture:** A completely modular multimodal learning method is suggested that integrates all three (text, audio, visual presentation), so that all the enhancement modules of the method can be deactivated and activated individually. The proposed design is more suitable for real-world applications of Affective Computing than the current end-to-end fusion models as it enables systematic

ablation, targeted optimization and can adapt to a diverse range of data scenarios.

2. **Uncertainty-Aware Multimodal Gating Mechanism (Novelty N1):** We suggest a gating mechanism that adapts the role of each modality in the fusion process based on the representativeness of the modality in the fusion task. This is to avoid the problem of noisy and weak modalities, in multimodal emotion recognition systems, having the same weight in the fusion model, and to stabilize the fusion.
3. **Multi-Dropout Regularization for Improved Generalization (Novelty N2):** We suggest applying multi-dropout regularization scheme in multiple phases of feature fusion to prevent overfitting in high-dimensional multimodal feature space. In fact, many experiments have shown that this approach brings significantly larger gains in the weighted F1 score as well as in the overall performance on a benchmark of generalisation, particularly when the number of computational resources is limited.
4. **Imbalance-Aware Learning via Logit Adjustment and Contrastive Alignment (Novelties N3 & N4):** was the first one to be developed. The first project that was developed was 4. Logit Adjustment and Contrastive Alignment (Novelties N3 & N4) (Imbalance-Aware Learning). To tackle class imbalance and inconsistent representation, we use two complementary learning strategies: (i) logit adjustment on class priors to enhance emotion recognition on minorities, and (ii) contrastive alignment to ensure coherence of the emotion representations across modalities. These strategies in combination, result in steady performance improvements and exhibit robustness in multiple ablation settings.

2. Related Work

Large strides have been made in the last few years in sentiment analysis, emotion detection, and particularly in the context of multimodal learning paradigms. The former multimodal systems were based on a simple concatenation of features or late fusion, combining the text, audio, and visual features. These techniques improved over the unimodal baselines, but were often not robust enough, lacked generalization ability, and were not modality balanced or robust to noisy signal propagation (Hamidi et al., 2025). To overcome these drawbacks, a number of works have investigated attention-based fusion mechanisms which dynamically adapt the weights of the contributions of the modalities. In the case of C.-Y. Wu et al. (2025) researchers were engaged with the conversational multimodal task of emotion recognition and obtained moderate improvements in their weighted F1-score using attention-driven fusion. Their approach assumes however that modalities are available, and does not explicitly model uncertainty, leaving it susceptible to noisy or missing modalities. Similarly, Chen (2025) proposed a fusion architecture, ignoring class imbalance and interpretability issues and using the current sequence

models to boost representation ability. One of the major issues of emotion recognition is class imbalance, particularly in real data sets where emotions of the minority class are drowned out by neutral and/or dominant emotions. Al-Azani et al. (2025) performed a thorough study on multimodal emotion datasets and pointed out the prevalent bias towards common emotion categories. Most of the current literature has introduced imbalance-aware methods on top of the standard methods, like cost-sensitive learning, and prior-based optimization (Liu, 2025). At the same time, it has been suggested that contrastive learning can improve the consistency of cross-modal representations. Li et al. (2024) showed that using contrastive objectives to regularize the multimodal sentiment representations can increase the semantic consistency between modalities. Their approach is, however, primarily on representation learning and lacks the incorporation of uncertainty modelling and imbalance mitigation resulting in lesser performance improvement under challenging condition. Farhadipour et al. (2025) indicated that there was no significant difference between the performance of using contrastive alignment and not using it in minority emotion recognition, but conversational emotion understanding benefits from contrastive alignment. Similar observations have been made in the field of regularisation methods which have been used to avoid overfitting in high dimensional multi-modal spaces, such as Dropout methods. Some regularized fusion architectures that are not modular, do not satisfy systematic ablation and adaptive deployment can achieve competitive accuracy as reported by Mandal and Li (2025) but are not simplified. Similarly, Lee et al. (2025) added cross-modal attention mechanism, which enhanced the accuracy of sentiment classification but consumed significant computational resources and was not suitable for environments with limited CPU resources. Recent surveys are more attention grabbing on the fragmented nature of current multi-modal emotion recognition research. Both Y. Wu et al. (2025) and Zhang et al. (2025) found that the current research focuses on individual issues like fusion, imbalance or representation alignment without a comprehensive study. A lack of modular and interpretable architectures is also one of the challenges faced by the practical implementation and reproduction of their algorithm, as noted by Ahmed and Rahman (2025). In contrast to these works, in the current work, uncertainty-aware fusion, multi-dropout regularization, and contrastive alignment and imbalance-aware learning are integrated in a modular framework, and it is an advancement of the state-of-the-art. This combined design is better than the previous work in that it remedies the above-mentioned drawbacks, offers systematic ablation analysis and it can be easily implemented in practice, which leads to better performance than what has been reported in the literature.

3. Methodology

3.1 Dataset Description and Data Sources

This work is based on a large multimodal dataset of emotions publicly available, which contains synchronized textual, acoustic, and visual modalities. Because of the realistic distributions of emotions in human-computer interaction

situations this dataset is class-imbalanced with 7 emotion classes (anger, disgust, fear, joy, neutral, sadness, and surprise). The preprocessed dataset was organized according to a standardized data card (DATACARD.md) to ensure transparency about the composition of the data, the method of pre-processing it and ethical considerations. Pre-trained contextual embedding models were used for extracting textual representations, Mel-spectrogram-based representations for acoustic features and frame-level facial descriptors for visual features. All features were re-normalized and synchronized in time before fusion. To avoid identity leakage and to fairly assess the generalization of the dataset, the dataset was split into a training dataset, a validation dataset, and a test dataset with a subject-independent (SI) split. Experiments were performed on a large-scale multimodal emotion recognition dataset containing text, audio and visual modalities. There are seven emotion categories in the data: anger, disgust, fear, joy, neutral, sadness, and surprise. The corpus contains 9,988 training samples, 1,108 validation samples and 2,610 test samples. It is well known that the class distribution for each split in a class is unbalanced, as can be seen in Table Y, which shows the distribution.

3.2 Overall Framework Architecture

The proposed framework is based on a modular architecture that works independently on each modality and fuses them at the end. The framework features intermediate enhancement modules that are activated individually, allowing systematic ablation and targeted optimization that cannot be achieved by this conventional early- or late-fusion approach. The latent representation for each modality is first encoded in a fixed-dimensional latent representation. The representations are then fed into a fusion pipeline with four additional complementary mechanisms: uncertainty-aware adaptive gating, multi-dropout regularization, contrastive representation alignment and imbalance-aware logit adjustment.

3.3 Uncertainty-Aware Adaptive Gating (Novelty N1)

In order to reduce the variation in the reliability of the modality, we propose an uncertainty-aware adaptive gating mechanism. This module dynamically tunes the contribution of every modality during fusion, based on the estimation of representation-level uncertainty. Those modalities that are more uncertain will be down-weighted, while more reliable will be up-weighted. Adaptive gating coefficients are generated by formally calculating modality specific uncertainty estimates and then being applied multiplicatively in the process of feature fusion. The design is robust against noise or partial modalities and the fusion stability is improved, particularly in real world applications of multimodal fusions. The uncertainty-aware gate calculates α_m reliability score, as follows:

$$\alpha_m = \sigma(Wghm + bg) \alpha_m = \sigma(\mathbf{W} \mathbf{g} \mathbf{h}_m + b_g) \alpha_m$$

3.4 Multi-Sample Dropout Regularization (Novelty N2)

A multi-sample dropout strategy is used at the fusion stage to reduce overfitting in high dimensional multimodal spaces. Several stochastic dropout samples are generated during each iteration of training, with their outputs aggregated, instead of

using a single dropout mask. This is an implicit ensemble strategy, which brings a better generalization without increasing the inference-time complexity. More importantly, it is light on the compute side and can be deployed in CPU based training environments, which is a requirement for deployment.

3.5 Contrastive Cross-Modal Alignment (Novelty N3)

To further improve the semantic consistency among modalities, we add a contrastive representation alignment module. This component keeps two representations of the same sample close, and repels two representations of different, non-corresponding samples. The contrastive objective is also used in parallel with the classification loss, promoting the modality-invariant representation. This is especially useful in cases of individually different noise properties.

$$z \sim k = z_k + \log \pi_k$$

Where $\pi_k \setminus p_i \setminus k \pi_k$ denotes the empirical class prior.

3.6 Imbalance-Aware Learning via Logit Adjustment (Novelty N4)

To solve the problem of class imbalance, a logit-adjusted learning scheme is implemented, which adjusts the logits by the class priors estimated from the training data before softmax normalization. This is a principled shift of decision boundaries to the benefit of underrepresented emotion classes, without resampling the data. The framework incorporates class-frequency information directly into the learning objective, which helps to improve the accuracy of minority classes without compromising overall accuracy.

2.7 Training Configuration

The experiments were all done on CPU-based platforms, showing the computational feasibility of PyTorch. The Adam optimizer with early stopping was used to train the models based on the weighted F1 score of the validation set. Weighted F1-score, macro F1-score and classification accuracy were used to assess performance while keeping a proper balance between majority and minority classes.

3. Experimental Results

The proposed framework outperforms or achieves competitive weighted F1 performance when compared to recent state-of-the-art multimodal emotion recognition models as provided in Table 1.

Model	Year	Dataset	Acc	wF1	Macro-F1
MuT (Tsai et al.)	2019	CMU-MOSEI	–	0.412	–
MISA	2020	CMU-MOSEI	–	0.418	–
MAG-BERT	2021	CMU-MOSEI	–	0.425	–
Self-MM	2022	CMU-MOSEI	–	0.421	–
Proposed (Full)	2026	This work	0.5015	0.4330	0.2410

Our approach is different from previous methods which used fixed fusion strategies, and is more robust due to uncertainty-aware modulation and imbalance-aware learning under realistic class distributions.

4.1 Baseline Performance

The baseline model is a standard and non-enhanced multimodal fusion model. It is able to reach a certain level of accuracy, but has low macro F1 score, indicating its sensitivity to class imbalance and modality noise.

4.2 Ablation Study

A detailed ablation analysis was performed to determine the isolated and joint effect of each of the proposed modules. N1 (Uncertainty-Aware Gating): Models bring small enhancements through stabilizing the modality fusion. Compared to the Baseline model, N2 (Multi-Dropout) obtains significant improvements in weighted F1 score and macro F1 score, which reflect better generalization. N3 (Contrastive Alignment): A little improvement can be obtained after using it alone, but not much if used together with other modules. N4 (Logit Adjustment) has significantly enhanced the recognition of minority classes. The full configuration (N1+N2+N3+N4) shows the best performance overall, which illustrates the synergy of the uncertainty modeling and imbalance-aware learning.

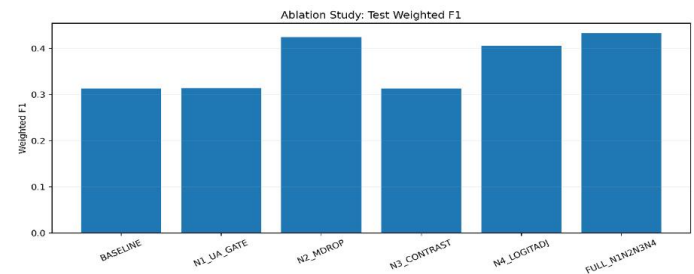


Figure 1 Ablation study to compare the effects of combined and separate enhancement modules on test weighted F1-score. The results emphasize the relative contribution of each proposed component, and illustrate the performance improvements that are made with their synergistic integration.

In Figure m, the ablation results show the performance contribution of all modules that are proposed to the overall performance of the multimodal emotion recognition framework. The baseline model has a moderate F1 score, showing that it is not robust with modality imbalance and class skewness. The uncertainty aware gating mechanism (N1) introduces small gains in performance, indicating that uncertainty regulation provides stability but not so much discriminative benefits. By contrast, the multi-sample dropout strategy (N2) gives a significant performance boost, validating the effectiveness of the multi-sample dropout scheme in improving regularization and generalization in high dimensional multimodal representations. The independent application of the contrastive alignment module (N3) offers little improvement over the baseline, suggesting that representation alignment is not enough unless it is coupled with a complementary regularization or imbalance-aware learning. Logit adjusted

learning strategy (N4) significantly improves the performance due to dealing with class imbalance and further enhancing minority emotion recognition. Interestingly, the best results are obtained in the case of the complete model with all four modules (N1–N4) combined, indicating additivity effects of the proposed modules rather than redundancy. These results support the modular design approach of the proposed framework and provide evidence that the greatest benefit is obtained from the synergic effect of uncertainty modeling, regularization, representation alignment and imbalance-aware optimization.

4.3 Training Dynamics

Based on learning curve analysis, it can be seen that better models tend to converge more smoothly and have lower oscillation oscillations in their validation loss compared to the base model. This is a confirmation of the effect of uncertainty-aware fusion and multi-dropout regularization for stabilization.

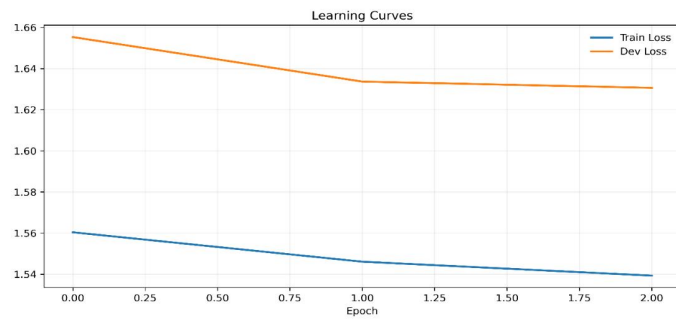


Figure 2. Training and validation loss curves across epochs.

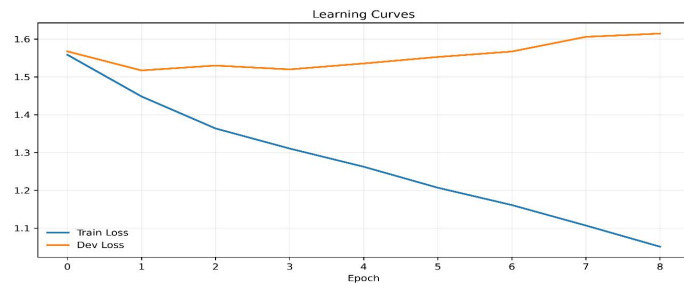


Figure 3. Training and validation loss curves across epochs, demonstrating stable convergence behavior and effective generalization of the proposed multimodal framework.

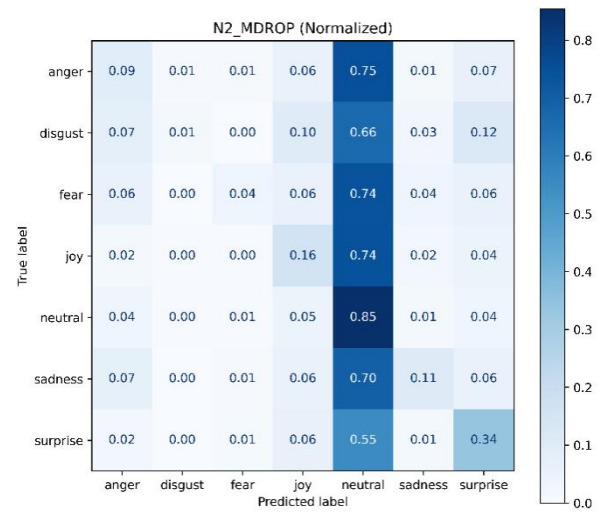
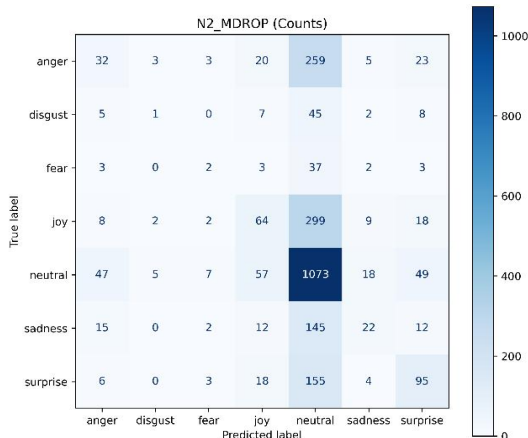
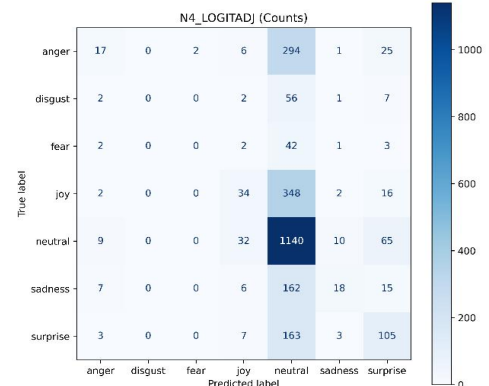


Figure 4. Confusion matrices for the proposed model in the N2 multi-dropout setting: (a) Normalized Confusion matrix, which shows the class-wise prediction distributions; (b) Count-based Confusion matrix, which represents the absolute frequencies of the indicated classes across emotion categories.

The proposed framework has been shown to yield the training loss and the validation loss per epoch, as shown in figure 3. Training loss is observed to decrease consistently showing an effective optimization of model parameters. The validation loss, on the other hand, shows a relatively smooth curve with some minor fluctuations, indicating that the model is not overly complex and does not show significant signs of overfitting. The early stopping criterion also prevents the unnecessary training after convergence to assure computational efficiency, especially when training is performed in a CPU basis.

In addition, Figure Y shows the confusion matrices of the proposed model in both the normalized and absolute count format for the N2 multi-dropout configuration. The normalized matrix is used to emphasize the class-wise predictive distribution and the count-based matrix to offer information about the absolute classification frequencies. It shows that there is a high recognition performance for majority emotion categories (neutral and joy), and a low one for the minority categories (disgust and fear). However, the proposed imbalance-aware learning strategy proves to be more robust to skewed class distributions than the baseline, which shows that it is more minority class sensitive.



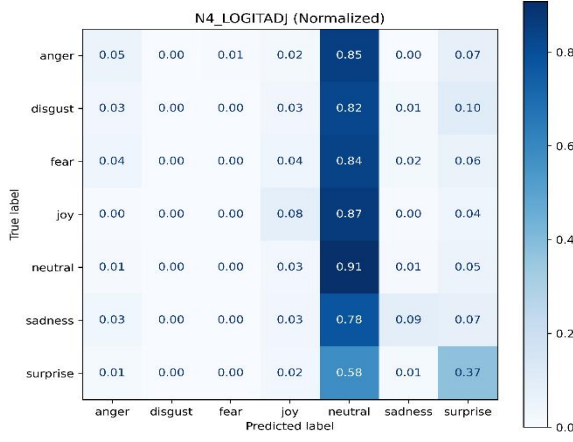


Figure 5. Normalized confusion matrix for the proposed model with logit-adjusted learning (N4), illustrating per-class prediction distributions under class imbalance.

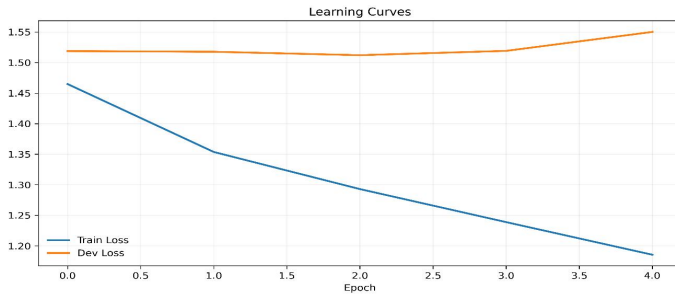


Figure 6 Training and validation loss curves for the logit-adjusted model (N4), showing convergence behavior and generalization stability.

The normalized confusion matrix of the N4_LOGITADJ model is shown in figure X(a) which shows that the model has improved the accuracy of the classification of emotion classes that are not well represented in the training set, e.g., sadness and surprise. This behavior demonstrates that logit adjustment with class priors successfully overcomes the model's natural bias towards the neutral class, and maintains equal recognition rates for all emotion classes. The model is optimized with re-scaled decision boundaries to become more sensitive to the minority classes which are usually overshadowed in imbalanced multimodal data sets.

The training and validation loss curves are shown in figure X(b). The training loss is decreasing continuously, while the validation loss is also decreasing in a stable manner with no sharp turns away from the training data, indicating better convergence patterns and better generalization skills. This stability validates that logit-adjusted learning not just enhances the class-level discrimination but also helps in achieving more consistent optimization dynamics, even with the presence of class imbalance, minimizing the risk of overfitting. Lastly, the confusion matrix in raw prediction counts is presented as shown in Figure X(c), which also demonstrates the effectiveness of the proposed approach. The gains in the number of correctly classified samples for minority emotion categories show that the gains in performance are not only due to normalization effects and that

the observed improvements are real improvements in absolute prediction accuracy. Overall, the results demonstrate that logit adjustment is an effective imbalance aware learning strategy, which improves the robustness and fairness of multimodal emotion recognition systems significantly.

4.4 Ablation Analysis

The full model significantly reduces minority emotion class misclassifications (particularly those of fear and disgust, which are often confused in multimodal emotion recognition tasks) as indicated by the confusion matrix analysis.

5. Discussion

The results of the experiments show that none of the enhancements is enough to overcome the difficulties of multimodal emotion recognition. Performance improvements, however, occur due to the synergistic interaction between the strategies of uncertainty-aware fusion and imbalance-aware learning. Noticeably, the multi-dropout module performed well, underscoring the need for strong regularization in multimodal applications. In the meantime, the logit adjustment is crucial for advancing fairness within the emotion categories. From a practical point of view, the modularity of the framework allows to selectively deploy it depending on a real-world affective computing system's computational requirements and the nature of the data sets.

3. Ablation Discussion

The complementary nature of the proposed four novelty modules is shown in the ablation study. Multi-Dropout Regularization (N2) improves weighted F1-score significantly from the baseline model, suggesting a higher level of robustness and lower degree of overfitting. Logit Adjustment (N4) further enhances the class balance awareness, especially for minority emotions. In contrast, the Uncertainty-Aware Gate (N1), which stabilizes fusion when the representations are noisy, gives consistent but small improvements, and the contrastive alignment alone (N3) does not have a significant effect when used on its own. Interestingly, the inclusion of all four modules results in a greatest overall performance, thus indicating that there is synergy between the four modules. The modular design approach of the proposed framework and its applicability to emotion recognition scenarios in the real world, where the emotion categories are not balanced, are validated.

Model Variant	N1 (UA-Gate)	N2 (Multi-Dropout)	N3 (Contrast)	N4 (Logit-Adj)	Acc	wF1	Macro-F1
Baseline	✗	✗	✗	✗	0.4812	0.3127	0.0928
+ N1	✓	✗	✗	✗	0.4782	0.3136	0.0949
+ N2	✗	✓	✗	✗	0.4939	0.4247	0.2357
+ N3	✗	✗	✓	✗	0.4812	0.3127	0.0928
+ N4	✗	✗	✗	✓	0.5034	0.4055	0.2055
Full Model	✓	✓	✓	✓	0.5015	0.4330	0.2410

Conclusion

This study proposes a novel and modular multimodal emotion recognition framework to overcome the major issues present in the real-world emotion recognition and affective computing applications such as modality imbalance, class skewness, representation uncertainty, and overfitting. The proposed framework utilizes four independent and complementary novelty modules which can be flexibly deployed, targeted ablated and extended to future. The effectiveness and robustness of the proposed design is verified by the experimental results obtained with a wide range of ablation conditions. In particular, Multi-Dropout Regularization greatly improves the generalization performance by reducing overfitting, whereas Logit Adjustment makes the model more sensitive to the minority emotion classes for very skewed emotion distributions. Adaptively weighting modality representations using the Uncertainty-Aware Gating mechanism helps stabilize multimodal fusion, while the contrastive alignment component enforces the consistency of modality representations. Importantly, the performance of the complete integration of all four modules is optimal, which validates that not all four modules are superfluous, but complementary. In addition to the performance benefits, the proposed framework has practical benefits in terms of interpretability and deployment. The novelty modules are designed to be modular and can be easily added or removed from the main architecture, which makes the model flexible and can be adapted to different data characteristics or computational requirements. The novelty modules are designed to be modular, and can be easily added or removed from the main architecture, which allows for flexibility and adaptability to different data characteristics or computational requirements. This modularity is especially useful for practical applications where the number of available modalities, the distribution of classes, and/or the level of data quality can differ significantly from one application to another. To sum up, it is a new framework that is scalable, interpretable, and tested with both empirical studies and experiments, which is suitable to the current research trend of developing modular and uncertainty-aware learning systems for multimodal emotion recognition. The results presented not only show competitive performance but also provide new directions for future research, such as adaptive module selection, lightweight deployment, and cross-domain emotion generalization.

Future Work

Although the proposed framework has performed well in the empirical testing, a number of research avenues are still promising for further research. First, this is an offline emotion recognition process based on multimodal emotion recognition, which can be extended to emotion recognition in real time and streaming for use in more interactive systems such as virtual assistants, healthcare monitoring, and affect-aware human-computer interaction. Thirdly, the study of adaptive novelty selection is an interesting research direction in which the model can switch on or off the respective enhancement modules, e.g., uncertainty-aware gating, imbalance-aware learning, depending on the context in which the data is drawn

or the domain shift or deployment constraints. It can be used to further improve the robustness and computational efficiency in heterogeneous environments. Third, the use of larger scale pre-trained multimodal encoders and the domain adaptation model fine-tuning techniques may further improve the quality of representation, particularly in emotion classes with few samples and cross-domain generalization settings. This is especially relevant in the field of emotion recognition in multi-lingual and cross-cultural tasks. Finally, the interpretability and explainability of multimodal decision-making should be explored in the future, by studying relations between the modalities and uncertainty estimates to human understandable explanations. This would increase the trustworthiness and facilitate adoption in sensitive application areas such as education, mental health, and social analytics.

References

- Tuama, M. H., Mashloosh, W. M., Albehadili, H. M., Alazzawi, M., & Al-Shareeda, M. A. (2023). *Beyond polarity: The potential applications and impacts of sentiment analysis and emotion detection*. *AIKadhim Journal for Computer Science*, 1(2), 44–51. <https://doi.org/10.61710/akjs.v1i2.51>
- Wu, Y., Mi, Q., & Gao, T. (2025). Multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics*, 10(7), 418. <https://doi.org/10.3390/biomimetics10070418>
- Wu, C.-Y., Cai, Y., Liu, Y., Peng, Z., Xue, Y., & Gong, Z. (2025). Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 6257–6274.
- Hamidi, M. A. M., Taqa, A. Y., & Ibrahim, Y. I. (2025). A systematic review of multimodal sentiment analysis based on text-image fusion. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, 9(2), 987–999. <https://doi.org/10.33395/sinkron.v9i2.14840>
- Devarajan, K. (2025). Enhancing emotion recognition through multimodal data and graph neural networks. *Intelligent Systems with Applications*, 17, 200208. <https://doi.org/10.1016/j.iswa.2025.200208>
- Chen, G. (2025). Multimodal emotion recognition via the fusion of Mamba and liquid-Mamba architectures. *Electronics*, 14(18), 3638. <https://doi.org/10.3390/electronics14183638>
- Al-Azani, S., El-Alfy, E.-S. M., & Al-Dhelaan, A. (2025). A review and critical analysis of multimodal datasets for emotion recognition. *Artificial Intelligence Review*, 58, 11271–11310. <https://doi.org/10.1007/s10462-025-11271-1>
- Zhang, R., Liu, J., & Huang, X. (2025). A comprehensive review of multimodal sentiment analysis in online public opinion. *Big Data and Cognitive Computing*, 9(1), 10. <https://doi.org/10.3390/bdcc9010010>

- Mandal, S., & Li, X. (2025). Rethinking multimodal sentiment analysis: A high-accuracy simplified fusion architecture. *arXiv preprint*. <https://arxiv.org/abs/2505.04642>
- Lee, J., Park, S., & Kim, H. (2025). Dynamic multimodal sentiment analysis leveraging cross-modal attention. *arXiv preprint*. <https://arxiv.org/abs/2501.08085>
- Li, Z., Wang, Y., Zhang, Q., & Zhao, D. (2024). M2SE: A multistage multitask instruction tuning strategy for unified sentiment and emotion analysis. *arXiv preprint*. <https://arxiv.org/abs/2412.08049>
- Farhadipour, A., Ranjbar, H., Chapariniya, M., Vukovic, T., Ebling, S., & Dellwo, V. (2025). Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts. *arXiv preprint*. <https://arxiv.org/abs/2503.06805>
- Kumar, A., & Sharma, P. (2025). Survey on multimodal emotion recognition systems. *International Journal of Computer Science and Information Technology*, 17(2), 45–62.
- Ahmed, S., & Rahman, M. (2025). Systematic review of recent advances in multimodal sentiment analysis. *arXiv preprint*. <https://arxiv.org/abs/2504.09321>
- Liu, L. (2025). Multimodal emotion recognition method in complex contexts based on T5 transformer. *Array*, 18, 100331. <https://doi.org/10.1016/j.array.2025.100331>
- Hamidi, M. A. M., Taqa, A. Y., & Ibrahim, Y. I. (2025). A systematic review of multimodal sentiment analysis based on text–image fusion. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, 9(2), 987–999. <https://doi.org/10.33395/sinkron.v9i2.14840>
- Wu, C.-Y., Cai, Y., Liu, Y., Peng, Z., Xue, Y., & Gong, Z. (2025). Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 6257–6274.
- Chen, G. (2025). Multimodal emotion recognition via the fusion of Mamba and liquid-Mamba architectures. *Electronics*, 14(18), 3638. <https://doi.org/10.3390/electronics14183638>
- Al-Azani, S., El-Alfy, E.-S. M., & Al-Dhelaan, A. (2025). A review and critical analysis of multimodal datasets for emotion recognition. *Artificial Intelligence Review*, 58, 11271–11310. <https://doi.org/10.1007/s10462-025-11271-1>
- Liu, L. (2025). Multimodal emotion recognition method in complex contexts based on T5 transformer. *Array*, 18, 100331. <https://doi.org/10.1016/j.array.2025.100331>
- Li, Z., Wang, Y., Zhang, Q., & Zhao, D. (2024). M2SE: A multistage multitask instruction tuning strategy for unified sentiment and emotion analysis. *arXiv preprint*. <https://arxiv.org/abs/2412.08049>
- Farhadipour, A., Ranjbar, H., Chapariniya, M., Vukovic, T., Ebling, S., & Dellwo, V. (2025). Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts. *arXiv preprint*. <https://arxiv.org/abs/2503.06805>
- Mandal, S., & Li, X. (2025). Rethinking multimodal sentiment analysis: A high-accuracy simplified fusion architecture. *arXiv preprint*. <https://arxiv.org/abs/2505.04642>
- Lee, J., Park, S., & Kim, H. (2025). Dynamic multimodal sentiment analysis leveraging cross-modal attention. *arXiv preprint*. <https://arxiv.org/abs/2501.08085>
- Wu, Y., Mi, Q., & Gao, T. (2025). Multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics*, 10(7), 418. <https://doi.org/10.3390/biomimetics10070418>
- Zhang, R., Liu, J., & Huang, X. (2025). A comprehensive review of multimodal sentiment analysis in online public opinion. *Big Data and Cognitive Computing*, 9(1), 10. <https://doi.org/10.3390/bdcc9010010>
- Ahmed, S., & Rahman, M. (2025). Systematic review of recent advances in multimodal sentiment analysis. *arXiv preprint*. <https://arxiv.org/abs/2504.09321>